THE UNIVERSITY
of EDINBURGH

Advanced Techniques for Subsurface Imaging

Bayesian Neural Networks and Marchenko Methods

Stephanie Earp

Thesis presented for the degree of
Doctor of Philosophy

THE UNIVERSITY
of EDINBURGH

School of Geosciences
2019

For Michael

Abstract

Estimation of material properties such as density and velocity of the Earth's subsurface are important in resource exploration, waste and CO_2 storage and for monitoring changes underground. These properties can be used to create structural images of the subsurface or for resource characterisation. Seismic data are often the main source of information from which these estimates are derived. However the complex nature of the Earth, limitations in data acquisition and in resolution of images, and various types of noise all mean that estimates of material parameters also come with a level of uncertainty. The physics relating these material parameters to recorded seismic data is usually non-linear, necessitating the use of Monte Carlo inversion methods to solve the estimation problem in a fully probabilistic sense. Such methods are computationally expensive which usually prohibits their use over areas with many data, or for subsurface models that involve many parameters. Furthermore multiple unknown material parameters can be jointly dependent on each datum so trade-offs between parameters deteriorate parameter estimates and increase uncertainty in the results.

In this thesis various types of neural networks are trained to provide probabilistic estimates of the subsurface velocity structure. A trained network can rapidly invert data in near real- time, much more rapidly than any traditional non-linear sampling method such as Monte Carlo. The thesis also shows how the density estimation problem can be reformulated to avoid direct trade-offs with velocity, by using a combination of seismic interferometry and Marchenko methods.

First this thesis shows how neural networks can provide a full probability density function describing the uncertainty in parameters of interest, by using a

form of network called a mixture density network. This type of network uses a weighted sum of kernel distributions (in our case Gaussians) to model the Bayesian posterior probability density function. The method is demonstrated by inverting localised phase velocity dispersion curves for shear-wave velocity profiles at the scale of a subsurface fluid reservoir, and is applied to field data from the North Sea. This work shows that when the data contain significant noise, including data uncertainties in the network gives more reliable mean velocity estimates.

Whilst the post-training inversion process is rapid using neural networks, the method to estimate localised phase velocities in the first place is significantly slower. Therefore a computationally cheap method is demonstrated that combines gradiometry to estimate phase velocities and mixture density networks to invert for subsurface velocity-depth structure, the whole process taking a matter of minutes. This opens the possibility of real-time monitoring using spatially dense surface seismic arrays.

For some monitoring situations a dense array is not available and gradiometry therefore cannot be applied to estimate phase velocities. In a third application this thesis uses mixture density networks to invert travel-time data for 2D localised velocity maps with associated uncertainty estimates. The importance of prior information in high dimensional inverse problems is also demonstrated.

A new method is then developed to estimate density in the subsurface using a formulation of seismic interferometry that contains a linear dependence of seismic data on subsurface density, avoiding the usual direct trade-off between density and velocity. When wavefields cannot be measured directly in the subsurface, the method requires the use of a technique called Marchenko redatuming that can estimate the Green's function from a virtual source or receiver inside a medium to the surface. This thesis shows that critical to implementing this work would be the development of more robust methods to scale the amplitude of Green's function estimates from Marchenko methods.

Finally the limitations of the methods presented in this thesis are discussed, as are suggestions for further research, and alternative applications for some of the methods. Overall this thesis proposes several new ways to monitor the subsurface efficiently using probabilistic machine learning techniques, discusses a novel way

to estimate subsurface density, and demonstrates the methods on a mixture of synthetic and field data.

Lay Summary

Just as doctors use techniques such as X-rays and MRI scans to see inside a patient's body, Geophysicists use seismic waves (waves of energy) to see inside the Earth. These waves help us to determine the different rocks and their properties. However, unlike in medicine where doctors can see the results almost instantaneously, the images created in geophysics often take weeks or months to produce. These investigations are important for Geophysicists to understand the structure of the subsurface if they want to monitor changes underground or find natural resources. Since in most cases it is impossible to gain direct access to the interior of the Earth many kilometres beneath the surface, these images are the only way for geophysicists to 'see' inside the Earth. Two of the rock properties geophysicists would like to know are the speed which a seismic wave travels through the rock and the density of that rock. This will help to determine the type of rock at the location in the Earth.

Geophysicists record either natural sources of energy such as earthquakes or environmental noise (called ambient noise) or energy created from artificial sources. These recordings, seismic data, are then converted into maps of subsurface properties through a process called *seismic tomography* that uses the fact that seismic waves travel through rocks in different ways depending on their composition and properties. Geophysicists can use properties derived from the seismic waves or the waves themselves to determine the rock parameters in a process called *inversion*.

The inversion process can be time consuming when uncertainty information on the result is required. In this thesis I use a method called *machine learning* to

perform the inversion faster. Machine learning is a technique that uses data to identify patterns and uses them to provide insights and predictions when given new, unseen data. The method used to create an algorithm that can identify patterns from data is called network training and takes a significant amount of time. However, once this 'training' has been completed results from new observations can be inverted rapidly. I use machine learning techniques to perform the geophysical inversion process that will give rapid seismic velocity results and the uncertainty of that result.

The estimation of density in the subsurface from seismic data is complicated as seismic waves are sensitive to multiple subsurface parameters. Whilst velocity can be independently determined by measuring the time it takes for a seismic wave to travel through the rock density often cannot be measured without the effect of other subsurface parameters, namely velocity. This means that when Geophysicists perform inversion of seismic data for density models they often need to have accurate estimates for velocity as well. Therefore the reliability of density results are dependent on the reliability of the velocity estimates. Ideally Geophysicists would like to find a new method for estimating density.

In this work I give an overview of machine learning methods and how they can be used to produce probabilistic estimates. I then present three situations where machine learning can be used to create 2D and 3D maps of seismic wave speed in the subsurface much faster than traditional methods. I discuss how they can be used to take snapshots at regular intervals to monitor changes in the subsurface. I also use a new method to create maps of density estimates from seismic data using a technique that reduces the dependence on other subsurface parameters.

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, either in whole or in part, in any previous application for a degree or professional qualification. Except where otherwise acknowledged, the work presented is entirely my own.

Stephanie Earp

2019

Acknowledgements

I would like to express my deepest gratitude to my supervisor Andrew Curtis for all his endless support, patience and advice over the last four years. I am grateful that he took a chance on me and gave me this opportunity. His energy and enthusiasm have encouraged me throughout the PhD and when times got tough he was there with motivational words and support. I will owe a lot of my future scientific success to his teaching and guidance.

I am also immensely grateful for my second supervisors, Giovanni Meles and Satyan Singh. I couldn't have asked for a more welcoming introduction to Edinburgh and to research than that given to me by Giovanni. His door was always open for me for helpful advice or discussions on my work and I appreciated his patience in allowing me to ask as many questions as I needed. His humerous lunchtime conversations were a particular highlight. The guidance and support Satyan showed me during the second half of my PhD was invaluable. His encouragement of my work and my ability as a scientist helped me to improve my research. I consider him a friend as well as a supervisor.

I would like to thank Reda Baina, Fredrik Hansteen, Ed Kragh, Alexander Kritski, Matteo Ravasi and James Rickett from the sponsors of the Edinburgh Interferometry Project for supporting my work. The regular sponsor meetings always created constructive comments and insightful discussions that improved my research and my scientific knowledge.

A big thank you goes to my office mates while I was at Edinburgh, Erica, Carlos, Atif, Angus, Xin, Dom, Lou and Hugo, who have made this journey much more enjoyable. Working in such a talented and supportive research group has

been an honour, I will miss the good company and scientific discussions. A special thank you goes to Claire for being a great friend over the last four years, my PhD experience wouldn't have been the same without you.

My friends outside of Edinburgh have been important in keeping me sane during my PhD, Chandra, Emily, Lizzie and Satish, thank you for always being at the end of the phone to listen to me complain or be ready in person with a glass of wine to celebrate my victories. Your friendships are invaluable.

A special thank you goes to my parents and sisters who have always supported me in my endeavors, wherever they take me in the world! Their unconditional support has given me the freedom to pursue my dreams.

Finally to my husband, Michael, you had no idea what you were getting yourself in to when you met me but I couldn't have completed this PhD without your support and patience. Thank you for always believing in me, even when I didn't believe in myself. I'm looking forward to the next chapter in our lives.

Table of contents

List of figures

List of tables

Nomenclature

Roman Symbols

G	Green's Function
n	Outward pointing normal
R	Reflectivity
t	Time
v	Velocity

Greek Symbols

μ	Mean
ω	Angular Frequency
ρ	Density
Σ	Covariance Matrix
σ	Standard Deviation

Acronyms

AVO	Amplitude variation with offset
FC	Fully connected
FWI	Full Waveform Inversion

GAN Generative Adversarial Network

GMM Gaussian Mixture Model

KL Kullback-Leibler

LRP Layer-wise relevance propagation

McMC Markov chain Monte Carlo

MDN Mixture Density Network

MLP Multilayer Perceptron

MSD Mean-squared difference

NADE Neural Autoregressive Density Estimator

NN Neural Networks

pdf Probability density function

PRM Permanent reservoir monitoring

RBM Restricted Boltzmann machine

SSIM Structural Similarity Index Measure

VAE Variational Autoencoder

Chapter 1

Introduction

Seismic waves are often used to image the interior of the Earth. Geophysicists study the response of the seismic waves that have travelled through the Earth to create maps of material parameters such as P- and S-wave velocity or density. By estimating these parameters they are able to infer the structure and rock composition of the interior of the Earth. Reliable parameter estimates are needed to create accurate structural images. In this thesis I explore novel imaging techniques that use seismic data to create estimates of velocity and density that can be used to gain knowledge about the subsurface.

1.1 Seismic Imaging

Seismic imaging or seismic tomography is the method used to estimate Earth parameters such as P- and S-wave velocity, density or anisotropy from seismic data. Tomography has been applied at a variety of length scales from global seismology (Aki et al., 1977; Woodhouse and Dziewonski, 1984; Trampert and Woodhouse, 1995; Shapiro and Ritzwoller, 2002) to resource exploration (Farra and Madariaga, 1988; Pratt et al., 1996; Hicks and Pratt, 2001; Bussat and Kugler, 2011) and near-surface studies (Xia et al., 1999; Socco et al., 2010; Mordret et al., 2013b). The parameters are used to determine geological structures and lithologies in the Earth's interior by displaying them as 2D or 3D maps displayed on a grid of pixels.

1.1.1 Imaging for Velocity

One of the first works in seismic tomography was Aki and Lee (1976) who used P wave first arrival travel times to invert for 3-dimensional velocity using data from local earthquakes. This was quickly extended to global tomography by Dziewonski et al. (1977) who imaged the Earth's mantle using P-wave travel time residuals. After these first studies many applications of travel time seismic tomography appeared such as crosswell tomography where active seismic sources in one well are used as a source of energy that is recorded on geophones in an adjacent well (McMechan, 1983; Bregman et al., 1989). The use of active source tomography increased rapidly because it was the only way to know the source location and timing accurately so that correct travel times could be calculated. The above examples used direct arrivals between a source and receivers. Bishop et al. (1985) first used seismic reflection data to determine a 2D velocity model and the reflector depth by minimising the difference between the travel times measured from the data and travel times generated from ray tracing through the model. With advances in compute power efforts were put into improving velocity images by jointly inverting travel times and amplitude information (Wang and Pratt, 1997; Wang et al., 2000) and more recently inverting the full waveform (Cary and Chapman, 1988; Pratt, 1999; Hicks and Pratt, 2001; Virieux and Operto, 2009; Operto et al., 2013).

Surface waves or long wavelength waves called normal modes can also be used to construct images of the Earth's interior. The velocity of surface waves depends on the shear-wave velocity structure with depth, and because different surface wave modes and frequencies oscillate to different depths they also travel at different speeds (Aki and Richards, 2002). Love and Rayleigh surface waves have been used for tomography at a global scale (Romanowicz, 1995; Trampert and Woodhouse, 1995; Shapiro and Ritzwoller, 2002; Meier et al., 2007a,b) and a regional scale in areas including Eurasia (Curtis et al., 1998; Ritzwoller and Levshin, 1998; Devilee et al., 1999; Villasenor et al., 2001), East Asia (Curtis and Woodhouse, 1997; Friederich, 2003), North America (Van der Lee and Nolet, 1997) and Australia (Simons et al., 1999; Fishwick et al., 2005).

The above applications rely on wavefields produced from an active source of energy such as an earthquake to be recorded on receivers encompassing the area of interest. An important step in surface wave tomography was the discovery that ambient background seismic noise could be transformed into useable signal by seismic interferometry (Lobkis and Weaver, 2001; Campillo and Paul, 2003; Snieder, 2004; Wapenaar, 2004; Curtis et al., 2006; Wapenaar and Fokkema, 2006). The cross-correlation of two recorded wavefields can produce an estimate of the Green's function between two receiver locations, as though one of the receiver locations was a point source and the other receiver records the response from that point source. This was important because seismologists no longer needed to record a point source of energy to perform tomography, but could create a previously unrecorded Green's function from passive sources such as oceanic waves, wind of anthropogenic activity. It also has the advantage that the Green's function source location was known exactly as it is the location of the receiver that is used as a point source.

Ambient noise tomography has been used widely to image the Earth's interior. First to apply this method was Shapiro et al. (2005) and Sabra et al. (2005) who inverted one month of ambient noise data from the US Array to invert Rayleigh group waves for high resolution images underneath California. Initially the most widespread approach was to perform a linearized inversion using Rayleigh wave group traveltimes (Sabra et al., 2005; Shapiro et al., 2005; Yao et al., 2006; Zheng et al., 2008), however the method has also been used to create Love wave velocity maps (Cho et al., 2007; Roux, 2009; Li et al., 2010). Bensen et al. (2008) recovered phase velocity maps in addition to group velocity maps for Rayleigh and Love waves across the United States and then extended this in Bensen et al. (2009) to invert for a 3D shear wave velocity structure. It has been used at a regional scale to invert for crustal structure in many locations such as underneath the British Isles (Nicolson et al., 2012, 2014), in the United States (Gerstoft et al., 2006; Lin et al., 2008) and Asia (Borah et al., 2014; Guo et al., 2015; Singer et al., 2017). Ambient noise tomography has also been used at a reservoir scale for estimation of shear wave velocity (Stewart, 2006; Bussat and Kugler, 2011; de Ridder and Dellinger, 2011; Mordret et al., 2013a) and seismic attenuation (Allmark et al.,

2018). Ambient noise tomography has enabled Geophysicists to image areas of the Earth's subsurface by turning what was previously considered as 'noise' into coherent signal. However, to create the coherent signal many hours or days of recordings need to be cross-correlated meaning that rapid inversion of data or real-time monitoring are not possible with this method.

With the advent of large and dense arrays wavefields can be recorded with a high temporal and spatial resolution so that the gradients of the wavefields can be interpreted using a wave equation in a technique often called seismic gradiometry. This means that it is possible to retrieve near-direct measurements of medium properties without interferometry. Curtis and Robertsson (2002) first showed that the derivatives of a wavefield recorded on volumetric arrays could be used to estimate P- and S-wave velocities using a 3D wave equation. The volumetric arrays were distributed in a 3D space so that spatial wavefield derivatives could be taken in depth as well as horizontally. Langston (2007a,b,c) showed that horizontal slowness could be determined from the inversion of first-order temporal and horizontal spatial derivatives, assuming that the wavefield contains only non-overlapping plane waves. This method can be used in 3D (Poppeliers et al., 2013) and has been applied to data from the US Array for phase velocity estimation (Liang and Langston, 2009; Liu and Holt, 2015) as well as in terrestrial and lunar near-surface studies (Edme and Yuan, 2016; Sollberger et al., 2016). de Ridder and Biondi (2015) extended the method of Curtis and Robertsson (2002) to horizontally propagating waves using the second order spatial and temporal gradients of the ambient seismic noise recordings to invert a 2D wave equation directly for phase velocity maps. Using just 10 minutes of recordings the method yielded results comparable to that of cross-correlational ambient noise tomography that needed to cross correlate over 40 hours of recordings to produce virtual seismic sources. The method was extended by Zhan et al. (2018) who combined it with compressive sensing to produce better results than standard tomography methods. The works of de Ridder and Biondi (2015) and Zhan et al. (2018) showed that it is possible to retrieve phase velocity maps using short length noise recordings that opens up the possibility of real-time monitoring. However this method is only applicable when dense arrays are available.

Many tomographic inverse problems are highly non-linear and non-unique, however to reduce computational expense the physics of the inverse problem is often linearized so that Geophysicists only solve for approximate solutions (Trampert and Woodhouse, 1995; Ritzwoller et al., 2002). Uncertainty estimates on the model solution are often desirable however it is difficult to estimate reliable uncertainties from linearized tomography (Shapiro and Ritzwoller, 2002; Bensen et al., 2009; Nicolson et al., 2012, 2014). Monte Carlo sampling algorithms can be used in Geophysical inverse problems to give probabilistic outputs defined by probability density functions (Mosegaard and Tarantola, 1995). Efforts have been made to solve the tomographic equations by Monte Carlo sampling within a Bayesian framework so that multiple tomographic models are provided that fit the given data and are consistent with given prior information (Sambridge and Mosegaard, 2002). In recent years tomographic problems have been solved using reversible jump Markov chain Monte Carlo (McMC) algorithms that address transdimensional problems so that the number of parameters in the inverse problem can be adjusted during the inversion process (Green, 1995; Green and Hastie, 2009). Bodin and Sambridge (2009) first used the reversible jump McMC algorithm in seismic tomography to produce a partially linearized sampling solution for 2D Rayleigh wave group velocity of the Australian continent. This was extended to a fully nonlinear inversion in 2D to estimate phase and group velocity maps (Galetti et al., 2015; Zheng et al., 2017) and to create 3D velocity models (Bodin et al., 2012; Hawkins and Sambridge, 2015; Piana Agostinetti et al., 2015; Galetti et al., 2017; Zhang et al., 2018). These methods involve repeatedly taking samples of potential velocity models and tens of millions of samples can be produced for high dimensional, complex problems. Therefore such solutions are computationally demanding, can require weeks of compute time and storage of large sample sets.

When rapid estimation of subsurface velocity models is required then dense arrays are needed to perform wave equation inversion or the physics of the inverse problem must be linearized. However wave equation inversion does not include uncertainty estimates on the final result and if non-linear physics is not included then the estimate is an approximation and it is difficult to obtain reliable uncertainties. If data sets are to be inverted rapidly without linearizing the

physics and also including full uncertainty information on the posterior result then alternative methods need to be used.

1.1.2 Imaging for Density

Density is an important parameter which helps to determine rock composition within the Earth. However it has not been well constrained using traditional seismic imaging techniques. This is largely because the sensitivity of seismic waves to density is mainly through impedance, the product of density and seismic velocity, which leads to inherent trade-offs between density and velocity structure in the seismic reflection and transmission responses (Kato and Kawakatsu, 2001). In addition density contrasts directly affect the amplitude of reflected and transmitted waves (Aki and Richards, 2002) but these amplitudes are difficult to measure directly or reliably due to instrumentation limitations such as ground coupling and local site effects (Berbellini et al., 2016).

At a global scale using only seismic data Dziewonski et al. (1975) and Dziewonski and Anderson (1981) estimated 1D averaged density models using normal modes and the Earth's total mass and moment of inertia. Normal modes display a sensitivity to gravity, however only for the lowest frequencies (Dahlen and Tromp, 1998) and therefore are most useful for global seismology. Tanimoto (1991) inverted long period seismic data jointly for density and S-wave velocity, they found that whilst S-wave resolution was good over all depths, the density resolution was good only at shallow depths. Most work over these longer length scales involves joint inversion of normal modes and gravity data to make the inversion more robust (Ishii and Tromp, 1999, 2001). More recently Blom et al. (2017) used full waveform inversion (FWI) to estimate density with low-frequency wavefields. Interestingly they conclude that using gravity data as additional information with FWI does not improve the inversion result.

Gravity data are directly related to density yet the solution to the inverse problem is inherently non-unique as they are potential-field measurements (Blakely, 1995). Thus strong prior information must be provided to constrain the inverse problem (Blakely, 1995), usually in the form of additional types of data such as

seismic data (Chaves and Ussami, 2013; Herceg et al., 2015). Root et al. (2016) compare gravity-based and seismic-based methods to estimate density over the British Isles. Gravity based density models have a higher lateral resolution compared to seismic-derived density. However, high resolution P-wave velocity models are needed to reduce the uncertainty in the gravity-derived density estimates.

In an exploration setting reflection data are used to estimate density with Amplitude variation with offset (AVO) inversion (Dębski and Tarantola, 1995; Downton and Lines, 2004) or estimated from P-wave velocity using Gardner's equation (Gardner et al., 1974). More recently density has been estimated using full waveform inversion. With the increase in compute power in recent years more elastic FWI has become feasible and efforts have been made to retrieve density (Jeong et al., 2012; Operto et al., 2013; Bai and Yingst, 2014; Li et al., 2019). Prieux et al. (2013) invert for density and P-wave velocity using a visco-acoustic FWI formulation that updates velocity and density in a hierarchical manner rather than updating both parameters simultaneously, this produces more stable inversion results. Jeong et al. (2012) use a two step method to invert for density through elastic FWI, first they invert for a velocity model using a constant density and then in a second step they re-invert the density model using the recovered velocity information. This method was shown to provide better density estimates than conventional waveform inversion methods. Despite these advances, inverting for density using waveform inversion schemes involves a trade-off with velocity or other elastic parameters and thus will deteriorate the final density result (Jeong et al., 2012).

Since accurate density models are important for determining rock composition it is necessary to obtain reliable density estimates. However current techniques to estimate density from seismic data alone such as AVO inversion or from Gardner's equation involve a trade-off with velocity. To improve density inversion results with seismic data, methods that do not rely on inverting for velocity jointly with density must be found.

1.2 Neural Networks

Neural networks are mathematical models that approximate a non-linear mapping between two parameter spaces. Their existence dates back to the 1940's the technique has gone through varying levels of popularity (Goodfellow et al., 2016). It was not until Rumelhart et al. (1985) published a paper that described an efficient method to train neural networks called *back propagation* that neural networks became popular and were applied to a wide field of problems. Back propagation uses a gradient descent approach and exploits the chain rule to propagate errors backwards through the network. Since then network size and structure have become larger and more complex with new types of network created to handle specific problems such as image recognition (LeCun et al., 1995), speech recognition (Hochreiter and Schmidhuber, 1997) or dimensionality reduction (Hinton and Salakhutdinov, 2006).

Neural networks can be used as generative models to describe the joint distribution between the parameter spaces. The two most common type of neural network generative models are Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). VAEs (Kingma and Welling, 2013) are based on autoencoder networks that copy the network input to the output via an internal layer that describes a lower dimensional representation of the input parameter space. A VAE can produce a generative model by adjusting the lower dimensional mapping in the autoencoder from a set of discrete variables to a set of probability distributions. The network can sample randomly from the probability distributions in the latent space to generate new models. They have been used in geostatistics to create a set of geological models that can be used for probabilistic inversion of hydrogeological data (Laloy et al., 2017).

Generative Adversarial Networks Goodfellow et al. (2014) are similar to VAEs in that they can generate samples from a probabilistic lower dimensional mapping. However, GANs consist of two networks, a generator and a discriminator network, that are trained at the same time to compete against each other. The generator network is trained to produce new samples that fits a given distribution, determined by the training set, while the discriminator network is trained to determine whether

data comes from the training set or is a newly generated sample. When the generator network has simulated the underlying distribution of the training set correctly the discriminator network will not be able to determine whether a sample is from the training set or the generator network. GANs have become increasingly popular as they can produce higher resolution images than VAEs (Dosovitskiy and Brox, 2016). They have been used in geostatistical modelling to produce realistic geological models that can be used for geostatistical inversion (Dupont et al., 2018; Mosser et al., 2017, 2018) as well as in seismic data processing for seismic noise attenuation and trace interpolation (Alwon, 2018).

Bishop (1994) proposed an adjustment to the traditional neural network so that it can provide an approximation to a probability distribution instead of a function output. This formulation is called a mixture density network (MDN) as the output of this modified neural network is a mixture density model such as a Gaussian Mixture Model (GMM). Instead of representing the distribution of variables in a latent space the MDN predicts the distribution directly on the output of the neural network model. MDNs have been used in a variety of applications such as predicting surf height (Carney et al., 2005), predicting financial time data (Ormoneit and Neuneier, 1996) and for self-driving car applications (Leung et al., 2016; Choi et al., 2018; Makansi et al., 2019). Graves (2013) used recurrent mixture density networks to generate realistic handwriting samples that could be adjusted for better legibility or modelled to a particular style of handwriting. This work inspired research into speech synthesis using mixture density networks (Zen and Senior, 2014; Wang et al., 2016) and the speech synthesis system in Apple's Siri in iOS11, Apple's voice assistant, was powered by a mixture density network (Capes et al., 2017).

In Geophysics neural networks have been applied extensively to many geophysical problems such as velocity analysis (Roth and Tarantola, 1994; Calderón-Macías et al., 2000; Araya-Polo et al., 2018), petrophysical analysis (Aristodemou et al., 2005; Maiti et al., 2007) and first break picking (Murat and Rudman, 1992; McCormack et al., 1993). For an introduction to neural networks in Geophysics and an overview of applications see Van der Baan and Jutten (2000), Poulton (2002) or Valentine and Kalnins (2016).

Devilee et al. (1999) first used neural networks in a probabilistic manner in Geophysics, to invert surface wave velocities for crustal structure by inverting for discretised marginal probability distributions. This was extended by Meier et al. (2007a,b) to use a mixture density network to give continuous probabilistic estimates of global crustal thickness and velocity structure. They demonstrated the importance of using noise in the data set as a regularisation in the neural network. Mixture density networks have also been used in non-linear petrophysical inversion of surface wave data for global temperature and water content (Meier et al., 2009), seismic data sets for subsurface porosity and clay content (Shahraeeni and Curtis, 2011; Shahraeeni et al., 2012) and well-log data for porosity and shale content (Niu et al., 2015). In earthquake seismology Käufl et al. (2014, 2015) show that mixture density networks can be used to rapidly invert waveform data for point-source parameters such as location and depth whilst incorporating realistic posterior uncertainties. They show that the method is robust to missing data and noise, therefore could be suited to early warning systems or real-time monitoring. Mixture density networks have also been used with statistical models to solve geophysical inverse problems with prior information (Nawaz and Curtis, 2017, 2018, 2019).

Mixture density networks have been shown to produce comparable results to Monte Carlo methods (Meier et al., 2007a; Shahraeeni and Curtis, 2011; Käufl et al., 2016) and can be used for rapid inversion. However the time needed to train a network means that for problems where the network will only be used once it is not an efficient method. The advantage of neural networks comes when a problem has to be solved repeatedly with new data. Since the network only needs to be trained once, future inversions are fast. This means that mixture density networks are ideally suited for subsurface monitoring problems where rapid non-linear inversions that include reliable uncertainty estimates are required.

1.3 Thesis Plan

In this thesis I explore novel methods to image for subsurface velocity and density parameters. In the case of velocity I use machine learning techniques to create

probabilistic velocity models by training networks that can be re-used for repeated inversions, avoiding the large computational expense of Monte Carlo sampling. For density I develop and apply a linear inversion scheme without a direct trade-off with velocity. What follows is a more detailed outline of the work in this thesis.

Chapter 2 provides an introduction to the theory and methods used in the Chapters 3, 4 and 5. I introduce Bayesian inverse problems and explain the concepts of marginal and joint probability distributions. Then I present the theory of neural networks and how they are used to emulate a mapping between two parameter spaces. The two structures that are used in this thesis are introduced: a multi layer perceptron and a convolutional network. Then I show how a neural network can be adapted to provide the probability distribution of an uncertain output with a type of network called a mixture density network. The chapter ends by explaining how each network is trained and the final probability density function is formed.

I then present three chapters applying neural networks to geophysical inverse problems to provide rapid, probabilistic solutions. First, in **Chapter 3** I present the mean and standard deviation of shear-wave velocity maps of the Grane field in the Norwegian North Sea obtained using mixture density networks. Networks are trained to invert phase velocities of Rayleigh-type Scholte surface waves for subsurface shear-wave velocity structures. I include uncertainties on the data input and show that this gives a more reliable mean velocity estimate when the data contain noise. The results are comparable to trans-dimensional Monte Carlo solutions yet are obtained far more rapidly. After training the networks shear-wave velocity maps are obtained in seconds and the networks could be applied to similar data types in monitoring situations.

Producing the inputs to the networks in Chapter 3 involves a prepossessing sequence that precludes this method if we wish near-instantaneous inversion results from field data. Therefore **Chapter 4** presents a method for near-real time monitoring with large and dense arrays by using a combination of gradiometry, wavefield inversion and mixture density networks. The method is tested using a small field data set: using only a short recording (minutes) the wavefield can be

inverted for phase velocity maps using gradiometry and wave equation inversion. I then invert these phase velocity maps point by point in a similar manner to Chapter 3 to give mean and standard deviation maps of velocity with depth, and I show that the results are again comparable in quality to Monte Carlo solutions.

The final neural network application in **Chapter 5** provides a non-linear, fully probabilistic inversion of seismic travel-times for 2D velocity maps. The methods are demonstrated on synthetic data sets and the effects of prior information on the results are discussed. I show that when realistic prior information is used, uncertainty loops can be seen in the standard deviation maps which are expected in non-linear inversions. Their presence therefore inspires confidence in the results. I also discuss how prior information can help to offset the effect of the curse of dimensionality to improve the mean velocity structure estimate.

Chapter 6 moves away from estimating subsurface velocity, to investigate a method to estimate subsurface density. I show, using synthetic data, that using wavefields recorded in the subsurface and a specific formulation of seismic interferometry, density can be estimated in the subsurface without the standard trade-off between density and velocity. Marchenko methods used to calculate wavefields at virtual receivers in the subsurface in cases where direct recordings in the subsurface are not present. These subsurface 'recordings' are theoretically sufficient to estimate density. Practical limitations of the Marchenko method render the density estimations less reliable as currently no method exists to provide true amplitude Marchenko Green's function estimates. I discuss a method to estimate the amplitudes in layered media and show how it improves the inversion results.

In **Chapter 7** I discuss the limitations of the neural network methods presented in this thesis and outline some ways that they could be improved. I discuss areas of research that could improve the reliability of neural network methods and propose options to advance progress of machine learning in Geophysics. I also discuss areas of current research that could improve the Marchenko estimates shown in Chapter 6 and suggest an alternative possible useful application for the

method. Finally **Chapter 8** summarises the main conclusions that can be drawn from the thesis.

1.4 Publications

The following publications have resulted from this thesis:

- Earp, S., A. Curtis, X. Zhang, and F. Hansteen, 2019, Probabilistic Neural Network Tomography across Grane Field (North Sea) from Surface Wave Dispersion data: arXiv preprint arXiv:1908:09588. This has been submitted to Geophysical Journal International. This publication is included in this thesis as **Chapter 3**. X. Zhang created the phase velocity maps and their uncertainties used in this work.

- Cao, R., S. Earp, S. A. de Ridder, A. Curtis, and E. Galetti, In Press, Near-Surface 3D Seismic Velocity models by Wavefield Gradiometry and Neural Network Inversion: Geophysics. Parts of this publication are included in this thesis as edited sections of **Chapter 4**. The other authors of the paper collected the field data and developed the gradiometry/wavefield inversion method, and produced the results from those methods. I used the results they derived from wavefield inversion to perform the mixture density inversion. My contribution to each section of the chapter is explained at the beginning of Chapter 4.

- Earp, S., and A. Curtis, 2019, Probabilistic Neural-Network based 2D Travel Time Tomography: arXiv preprint arXiv:1907.00541. This has been submitted to Neural Computing and Applications. This publication is included in this thesis as **Chapter 5**.

The following conference abstracts have also resulted from this thesis:

- Earp, S., and A. Curtis, 2019a, Neural network Travel-Time tomography: Presented at the 81st EAGE Conference and Exhibition 2019 Workshop Programme.
- Curtis, A., R. Cao, S. Earp, X. Zhang, S. De Ridder, and E. Galetti, 2019, Near-real time 3D seismic velocity and uncertainty models from Ambient Noise, Gradiometry and Neural Network Inversion: Presented at the 81st EAGE Conference and Exhibition 2019 Workshop Programme.

Chapter 2

Neural Network Inversion Method

This Chapter presents the concepts of neural networks and the methods I use in the following chapters of this thesis. First, I present the solution of the inverse problem in a Bayesian framework and introduce the concept of joint and marginal posterior probabilities. Then I discuss the concept of neural networks for geophysical inversion and show how these can be trained to produce posterior probability density functions.

2.1 Bayesian Inverse problems

We want to solve inverse problems in a probabilistic framework to find the posterior distribution of Earth models $\mathbf{m}$ that fit some given data $\mathbf{d}$, written as $p(\mathbf{m} \mid \mathbf{d})$. This is defined as (Tarantola, 2005)

$$p(\mathbf{m} \mid \mathbf{d}) = k\, p(\mathbf{d} \mid \mathbf{m})\, p(\mathbf{m}) \tag{2.1}$$

where $p(\mathbf{m})$ represents the prior probability density on the parameter space, $p(\mathbf{d} \mid \mathbf{m})$ represents the conditional probability of some data given the specific parameters (known as the likelihood) and k is a normalisation constant. The likelihood can be seen as a measure of how well a parameter $\mathbf{m}$ explains the data $\mathbf{d}$. Equation 2.1 shows how prior information about the parameters are updated by the data to produce the posterior distribution. If the data do not contain

any new information about the parameters we cannot learn anything and the posterior will equal the prior. Since inferences of the posterior are relative to the prior, Bayesian solutions to inverse problems depend on prior assumptions. In Chapters 3 and 4 the prior is selected based on expected geological information of the area whilst in Chapter 5 the effect of prior information on the inversion results is tested.

The posterior pdf $p(\mathbf{m} \mid \mathbf{d})$ represents a description of the full posterior uncertainties, not only the variance of each parameter but also the dependencies between parameters. These dependencies can create trade-offs between parameters for example if the velocity at one depth increased then the velocity below that must decrease in order to still fit the data. In multidimensional problems (where the dimensionality of $\mathbf{m}$ is greater than 1) we often need to make inferences about a set of individual parameters enumerated with index i, and hence must calculate the marginal posterior distribution $p(m^i \mid \mathbf{d})$. This can be obtained by integrating over all parameters j that are not of interest

$$p(m^i \mid \mathbf{d}) = \int_{m_k, k \neq i} p(\mathbf{m} \mid \mathbf{d})\, dm_k \tag{2.2}$$

If trade-offs between two parameters selected from a larger multidimensional problem are required then Equation 2.2 can be extended to give the joint marginal posterior,

$$p(m^i, m^j \mid \mathbf{d}) = \int_{m_k, i \neq k \neq j} p(\mathbf{m} \mid \mathbf{d})\, dm_k \tag{2.3}$$

This can be calculated by inverting for both parameters together by joint inversion, or alternatively the joint marginal posterior can be created from the product of conditional $p(m^j \mid m^i, \mathbf{d})$ and 1D marginal pdfs $p(m^i \mid \mathbf{d})$ (Tarantola, 2005)

$$p(m^i, m^j \mid \mathbf{d}) = p(m^j \mid m^i, \mathbf{d}) \times p(m^i \mid \mathbf{d}) \tag{2.4}$$

Trade-off relations can involve any number of parameters but in the following chapters I focus on estimating 1D and 2D posterior marginal distributions $p(m^i \mid \mathbf{d})$ (Equation 2.2) and $p(m^i, m^j \mid \mathbf{d})$ (Equation 2.4), respectively.

2.2 Neural Networks

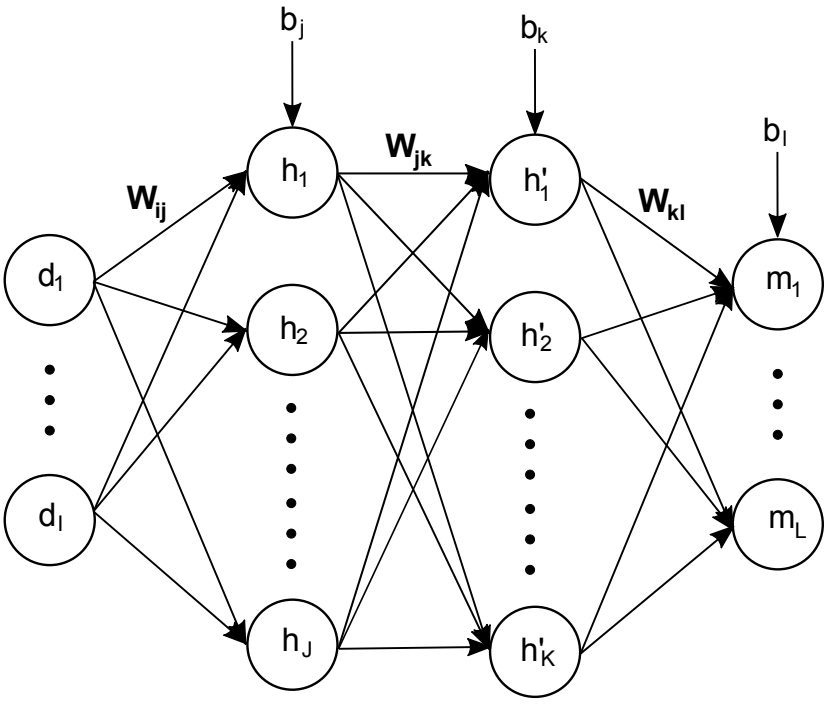

Figure 2.1 A schematic of a feed-forward Multilayer Perceptron with I input units, two hidden layers with J and K units respectively, and L output units. The two sets of trainable parameters are the weights $\mathbf{W}$ and biases $\mathbf{b}$ which are associated with the continuous arrows which represent the flow of information (calculated values).

A neural network (NN) is essentially a mathematical relationship that defines a mapping between two parameter spaces. If a forward problem is well-defined or easily calculated but the inverse problem is more complex or difficult to calculate (as in many geophysical inverse problems) neural networks can be useful for mapping the inverse relationship. The simplest form of a NN is called a feed-forward Multilayer Perceptron (MLP) where information is fed forward through the network from input to output, Figure 2.1. I use this form of neural network in Chapters 3, 4 and 5. At each layer the inputs are weighted and summed before being passed through an activation function $g(\cdot)$ to provide an output value that becomes the input for one unit in the following layer, by an equation in the form

$$h_j = g(b_j + \sum_{i=1}^{I} W_{ij} d_i) \tag{2.5}$$

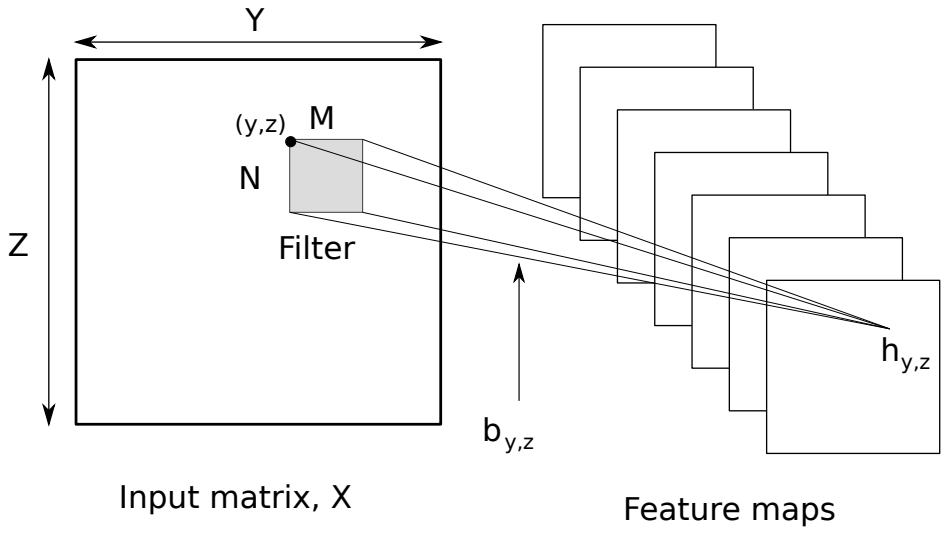

Figure 2.2 A schematic of a Convolution layer with an input matrix of size (Y, Z) . The filter F is of size (M, N) located at point (x, y) in the model. The trainable values in the filter are multiplied by the corresponding values in the input and summed along with the bias $b_(y, z)$ to give the point $h_{y,z}$ in the feature maps. Multiple feature maps are created using different values in the filter F.

Here d_i are the inputs in the input layer, h_j is the output value that is passed to the following layer, and W_{ij} and b_j are constants called the layer weights and biases respectively. The activation function can be chosen to create non-linearity in the mapping which allows us to use the network to emulate complex functions by varying $\mathbf{W}$ and $\mathbf{b}$ appropriately.

Convolutional networks are a different kind of neural network that use a convolution instead of general matrix multiplication for at least one of the layers. They are designed to be used with data that has a temporal or spatial correlation between neighbouring data points such as time series or image data (Goodfellow et al., 2016). In a convolutional layer a filter is passed over the input data and at each location the dot product between the filter-sized patch and the filter is computed and passed to the next layer 2.2. In a 2D case, for an input matrix $\mathbf{X}$ of size (Y, Z) and a filter $\mathbf{F}$ of size (M, N) the output from a convolutional layer h would be

$$h_{y,z} = g\left(b + \sum_m \sum_n F_{m,n} X_{y+m,z+n}\right) \tag{2.6}$$

where $b_{y,z}$ is the value of the bias constant. The filter is a set of weights used to detect *features* in the data. It is smaller than the size of the input matrix so that it can be passed over the matrix around multiple points to detect the same feature anywhere in the data (Goodfellow et al., 2016). For this reason the output layer is often called a *feature map*. The filter is moved across the input matrix at a set step-size in pixel location, referred to as *stride*: usually in two dimensions this is $(1, 1)$ and this is what we use in this work. Multiple filters are used at each layer so that many features can be extracted in parallel.

Parameters $\boldsymbol{W}$ and b in the MLP and the $\boldsymbol{F}$ and b in a convolutional network are tuned by *training* the network. This process adjusts the mapping to emulate any desired function provided an appropriate choice is made for the activation function $g(\cdot)$. The network is trained on a set of T input-output pairs $D = \{\mathbf{d}^t, \mathbf{m}^t\}$ where $\mathbf{d}$ is the network input and $\mathbf{m}$ is the desired network output. During training the parameters are tuned so that the network can accurately map the input to the output function. It does this by varying $\mathbf{W}$ and $\mathbf{b}$ so as to minimise a cost function E which measures the difference between the network output and the true output

$$E_{MLP} = \frac{1}{2}\sum_{t=1}^{T}(y^t - m^t)^2 \tag{2.7}$$

where $\mathbf{y}$ is the network output (which depends on $\mathbf{W}$ and $\mathbf{b}$) and $\mathbf{m}$ is the true output. In this work I train using standard back propagation (Rumelhart et al., 1985), with various improvements outlined as I use them below. Once trained, the network can be applied to new input data to give an estimate of unknown models. The input and output data can be any size and contain any data that can be represented numerically. The NN learns the mapping between input and output pairs and so as long as the inference data set matches the training data set in in parameter size and type, the NN will be able to estimate a model.

2.3 Mixture Density Network

An MLP trained by minimizing the sum-of-squared errors E_{MLP} will output an approximation to the conditional mean of a probability distribution $p(\mathbf{m} \mid \mathbf{d})$

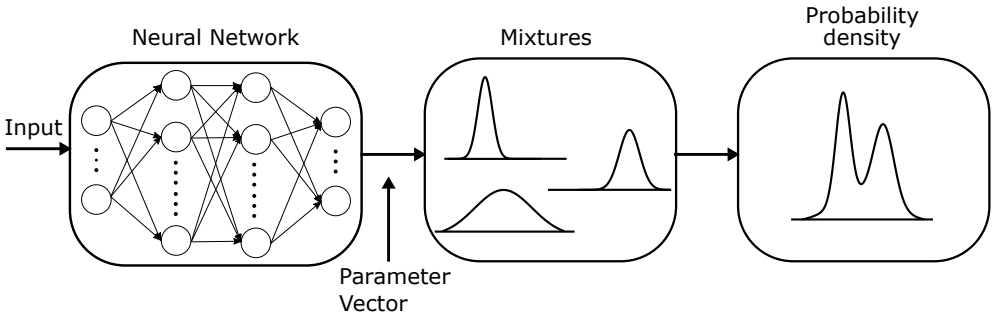

Figure 2.3 A schema of a Mixture Density Network as described by Bishop (1995). The output of the Neural Network gives a parameter vector that defines the Gaussian kernel for each mixture of Gaussians. These are then summed to give the probability of the model $p(\mathbf{m} \mid \mathbf{d})$.

(Bishop, 1995). A class of neural networks called mixture density networks provide a framework for modeling probability distributions in a similar way to how the MLPs can approximate function outputs (Figure 2.3). They are trained on the same pairs of data as an MLP but instead of providing an estimate of the model they provide an approximation to the probability distribution $p(\mathbf{m} \mid \mathbf{d})$, given by a mixture of Gaussian kernels

$$p(\mathbf{m} \mid \mathbf{d}) = \sum_{i=1}^{M} \alpha_i(\mathbf{d}) \Theta_i(\mathbf{m}|\mathbf{d}) \tag{2.8}$$

where α_i is the mixture or amplitude parameter that attaches relative importance to each Gaussian kernel, M is the number of Gaussians in the mixture, and Θ_i are Gaussian density functions given by

$$\Theta_i(\mathbf{m} \mid \mathbf{d}) = \frac{1}{\prod_{k=1}^{c}(\sqrt{2\pi}\sigma_{ik}(\mathbf{d}))} exp\left\{-\frac{1}{2}\sum_{k=1}^{c}\frac{(m_k - \mu_{ik}(\mathbf{d}))^2}{\sigma_{ik}^2(\mathbf{d})}\right\} \tag{2.9}$$

where c is the dimensionality of $\mathbf{m}$. The mean of the Gaussian kernel is described by $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{ic})$ and the covariance is assumed to be a diagonal matrix $\boldsymbol{\Sigma} = diag(\sigma_{i1}, ..., \sigma_{ic})$. The mixture parameter α_i, the means $\boldsymbol{\mu}_i$ and the standard deviations $\boldsymbol{\sigma}_i$ fully define the Gaussian kernels and hence the output of the MDN. Previous works use an isotropic Gaussian with diagonal covariance matrix and equal diagonal elements (Meier et al., 2007b,a; De Wit et al., 2013; Käufl et al., 2016) to estimate the full posterior pdf, however I follow the theory of Shahraeeni

and Curtis (2011) and use a diagonal covariance matrix with unequal diagonal elements as it is then usually possible to estimate the pdfs with far fewer kernels , and fewer network paramters overall despite the increase in parameters per kernel (Shahraeeni and Curtis, 2011). We could approximate more complex Gaussian kernels with a full covariance matrix as in Williams (1996), however this is more computationally expensive and the above method appears to be sufficiently flexible for the scope of this work. In Chapters 3 and 4 the dimensionality of $\mathbf{m}$ in each network is 1 and therefore the method outlined above is equivalent to using an isotropic Gaussian. However in Chapter 5 we also train networks where the dimensionality of $\mathbf{m} > 1$ and in this case we use the full formulation in Equation 2.9.

The number of kernels in the mixture dictates the complexity of the final probability distribution, and the number of network outputs is given by $(2c+1)\times M$ compared with the output of a standard MLP that has c outputs. The raw parameters output from the MDN z_i^α, z_{ik}^σ and z_{ik}^μ need to be transformed into usable parameters for Equations 2.8 and 2.9. The amplitude parameters α_i are required to sum to 1 and are calculated through the *softmax* function

$$\alpha_i = \frac{exp(z_i^\alpha)}{\sum_{j=1}^{M} \exp(z_j^\alpha)} \tag{2.10}$$

whilst the standard deviations have to be positive by definition, so we apply the transformation

$$\sigma_{ik} = \exp(z_{ik}^\sigma) \tag{2.11}$$

The means can take any real value so can be used directly such that

$$\mu_{ik} = z_{ik}^\mu \tag{2.12}$$

Training an MDN requires that we have a way to predict appropriate values for these parameters given any input data: in this thesis I do so using a conventional MLP or a convolutional neural network, which shows that an MDN typically has a standard neural network at its core. During training the network is adjusted to maximize the likelihood of the desired probability density function given the

training data. This can be achieved by minimizing the negative log likelihood function (Bishop, 1995)

$$E_{MDN} = -\sum_{n=1}^{N} ln \left\{ \sum_{i=1}^{M} \alpha_i(\mathbf{d}) \Theta_i(\mathbf{m}_n|\mathbf{d}_n) \right\} \tag{2.13}$$

Once the network has been trained, the outputs can be used to generate the predicted posterior probability distributions using Equations 2.8, 2.9. 2.12, 2.11 and 2.12. This gives a more complete description of the family of models that are consistent with the data compared to the output of a standard MLP. In this thesis $\mathbf{d}$ represents seismic data related parameters and $\mathbf{m}$ are Earth parameters, the precise composition of the inputs and outputs are explained in each chapter.

2.4 Network Training

Mixture density network training corresponds to the minimisation of the cost function 2.13. It is performed using gradient-based optimisation of the network's internal parameters called Adaptive Moment Estimation (Kingma and Ba, 2014). The optimiser computes adaptive learning rates for each parameter based on the average first and second moments of the gradients. Gradient-based methods typically operate iteratively and are therefore sensitive to the initialisation of the initial parameters. I do this using the Glorot uniform initialiser (Glorot and Bengio, 2010), which draws samples from a uniform distribution within $[-D, D]$ where

$$D = \sqrt{\frac{6}{x_{in} + x_{out}}} \tag{2.14}$$

where x_{in} is the number of nodes in the input to the layer and x_{out} is the number of nodes in the output of the layer.

To stabilise the training process and improve convergence the input vectors are pre-processed before network training (Bishop, 1995). I standardise the input data $\boldsymbol{d}$ to have zero mean and variance of 1 so that

$$\hat{d}_i = \frac{d_i - \bar{d}_i}{\sqrt{\sigma_{di}}} \tag{2.15}$$

where $\bar{d}_i$ and σ_d are the vector mean and variance of the data set respectively. This gives an equal weighting for the input vector components. The transformation is calculated on the training set and the same transformation is then applied to the test set and any new data used by the network. Any synthetic noise is added to the data before standardisation.

The aim of training is to learn a mapping between some input and output parameter spaces that can be applied to unseen data and produce accurate results. To achieve this the network should be able to generalise so that it can make accurate predictions on data it has never seen before. If the network is relatively simple and contains few trainable parameters it will create a smooth mapping which will not encapsulate the complexities of the underlying function. The network will be under-fitted leading to large errors, bias, in network predictions. Conversely, a complex network with many trainable parameters will predict the training data exactly but fail to generalise. When presented with previously unseen data the network will display high variance and make inaccurate predictions, the network will be *over-fitted*. This balance over the complexity of network models is known as the *Bias-Variance trade off*. Both scenarios need to be avoided when training a network.

Two methods are employed to mitigate the effects of a poorly fitted network. Firstly, Gaussian noise is added to the training data $\mathbf{d}$: this forces the network to find a smooth mapping and avoids the network fitting the details of the training data (Bishop, 1995). Furthermore a validation set is used during training to monitor the error (Equation 2.13) as well as the error on the training set. Generally, the error on the validation set will decrease until the network starts to over-fit the data, when it will begin to increase. When the validation error stops decreasing training is halted and the set of weights that minimize the validation error are used for the final network. A separate test set is then used to determine the accuracy of the network on unseen data since the training and validation set are used during training. If the error on the test set is much higher than the training and validation set the network will be over-fitted and the complexity of the network must be changed and the network re-trained.

Since the training method is iterative, the trained NN's are sensitive to the random parameter initialization and to the network configuration (internal structure). We train an ensemble of multiple networks with different configurations and combine them to give a group of networks - a so-called *mixture of experts*. In theory networks trained independently may make good predictions for different reasons and under different inputs (in our case, data vectors); using a combination of networks therefore often results in better generalisation of performance to unseen data and improves prediction accuracy (Dietterich, 2000). Bishop (1995) shows that the upper bound on the ensemble error is given by the average of the errors of the individual networks. We construct the ensemble by a weighted average of network outputs, where each weight is determined by the performance of the associated network on the test data set (or simply *test set*). The posterior probability distribution is thus estimated by

$$p(\mathbf{m} \mid \mathbf{d}) = \sum_{i=1}^{M} \sum_{l=1}^{L} \frac{\eta_i \alpha_{il}}{\sum_{k=1}^{M} \eta_k} (\mathbf{d}) \Theta_{il}(\mathbf{m}|\mathbf{d}) \tag{2.16}$$

where

$$\eta_i = exp(-E_{MDN,i}) \tag{2.17}$$

The final estimate of probability distribution $p(\mathbf{m} \mid \mathbf{d})$ contains LM Gaussian kernels. In Chapters 3 and 5 the results are from an ensemble of MDNs. In Chapter 4 multiple MDNs are trained and the one which performs best on the test set is used for the final results.

Chapter 3

Probabilistic Neural Network Tomography beneath Grane field (North Sea) using surface wave dispersion data

Surface wave tomography uses measured dispersion properties of surface waves to infer the spatial distribution of subsurface properties such as shear-wave velocities. These properties can be estimated vertically below any geographical location at which surface wave dispersion data are available. As the inversion is significantly non-linear, Monte Carlo methods are often used to invert dispersion curves for shear-wave velocity profiles with depth to give a probabilistic solution. Such methods provide uncertainty information but are computationally expensive. Neural network based inversion provides a more efficient way to obtain probabilistic solutions when those solutions are required beneath many geographical locations. Unlike Monte Carlo methods, once a network has been trained it can be applied rapidly to perform any number of inversions. A class of neural networks called mixture density networks are trained to invert dispersion curves for shear-wave velocity models and their non-linearised uncertainty. Mixture density networks are able to produce fully probabilistic solutions in the form of weighted sums of multivariate analytic kernels such as Gaussians, and I show that including data uncertainties in the mixture density network gives more reliable mean velocity estimates when data contains significant noise. The networks were applied to data from the Grane field in the Norwegian North sea to produce shear-wave velocity

maps at several depth levels. Post-training probabilistic velocity profiles were obtained with depth beneath 26,772 locations to produce a 3D velocity model in 21 seconds on a standard desktop computer. This method is therefore ideally suited for rapid, repeated 3D subsurface imaging and monitoring.

3.1 Introduction

Seismic surface waves travel around the surface of the Earth but are sensitive to heterogeneity in elastic properties within the subsurface. Different frequencies of surface waves travel at different speeds since they depend mainly on the shear-wave velocity structure at different depths. Surface wave tomography uses this property (called dispersion) to infer the spatial distribution of subsurface shear velocities over global scales (Woodhouse and Dziewonski, 1984; Trampert and Woodhouse, 1995; Shapiro and Ritzwoller, 2002; Zhou et al., 2006; Meier et al., 2007a,b), regional scales (Montagner and Jobert, 1988; Curtis and Woodhouse, 1997; Curtis et al., 1998; Ritzwoller and Levshin, 1998; Devilee et al., 1999; Villasenor et al., 2001; Simons et al., 2002) and reservoir scales (Bussat and Kugler, 2011; de Ridder and Dellinger, 2011; Mordret et al., 2014).

Surface wave tomography is often performed using a 2-step inversion scheme (Trampert and Woodhouse, 1995; Ritzwoller et al., 2002). In step 1, travel times of surface waves between pairs of known locations are measured at various fixed periods, then used to create geographical phase or group velocity maps at each period using 2D tomography. In step 2, the dispersion properties (speed of the waves at different periods – often referred to as a dispersion curve) at each point on the 2D map are then inverted to estimate a 1D shear-wave velocity profile with depth below that point. The 1D velocity profiles beneath many geographical locations can then be placed side-by-side and interpolated to create a 3D model of the subsurface.

Both of the 2-step surface wave inverse problems are non-linear. They can be solved approximately by partially linearized (Bodin and Sambridge, 2009), or fully non-linear (Rawlinson et al., 2014; Galetti et al., 2015, 2017) Monte Carlo methods.

These types of approaches provide relatively robust estimates of the range of possible shear wave velocity structures with depth that are consistent with the measured surface wave speeds (often referred to as the solution *uncertainty*) by using the Markov chain Monte Carlo (McMC) algorithm to perform the inversions in a Bayesian framework. However, all existing sampling based methods, including the direct (1-step) 3D Monte Carlo tomography method of Zhang et al. (2018), are extremely demanding computationally. If large data sets are to be inverted rapidly while maintaining the ability to assess post-inversion uncertainties without making linearizing approximations to the Physics, different methods are needed to speed up fully non-linear inversions.

I take an alternative approach and use neural networks to perform non-linear inversion of the phase velocities of Rayleigh-type Scholte surface waves (referred to simply as Rayleigh waves) for subsurface shear-wave velocity over length scales ~1-10km. Neural networks (NN's) approximate a non-linear mapping between two parameter spaces. The mapping is inferred from a set of examples of inputs and corresponding outputs of the real mapping (these examples are called *training* data). Using certain types of NN-based methods, uncertainties in the mapping can be output by the network. NN's are therefore useful for problems where the forward mapping is well known or simple to calculate (in order to construct many training data) but the inverse mapping is complex or costly to determine directly. In such cases training data can be generated by applying the forward mapping to many sets of model parameter values, after which the NN can be trained to map in the inverse direction, taking the measurable data as input and outputting model parameter estimates.

Once trained, NN's can be applied to calculate the mapping for any input parameters extremely efficiently. For this reason neural networks have become increasingly popular for solving geophysical problems in recent years. Applications include well-log analysis (Aristodemou et al., 2005; Maiti et al., 2007), first arrival picking (Murat and Rudman, 1992; McCormack et al., 1993), fault detection (Araya-Polo et al., 2017; Huang et al., 2017) and velocity analysis (Roth and Tarantola, 1994; Calderón-Macías et al., 2000; Araya-Polo et al., 2018). However all of these methods provide only deterministic estimates of the inverse problem

solution (in most cases, the mean model estimate). Neural networks can also be used in a Bayesian sense to give fully probabilistic solutions. They were first used in Geophysics to estimate Bayesian uncertainties by Devilee et al. (1999) who inverted surface wave phase and group velocities for large-scale subsurface velocity structure and interface depths. They inverted regional dispersion curves for discretised probability distributions of crustal thickness across Eurasia using histogram and median networks, and analysed the trade-off between crustal thickness and velocity structure. Meier et al. (2007a,b) improved this method by using mixture density networks (MDN's) to give continuous probabilistic estimates of global crustal thickness and crustal velocity structure. A MDN is a type of network that maps an input vector to a probability density function (pdf) rather than to a single set of output values (Bishop, 1995). Since the work of Meier et al. (2007a,b), mixture density networks have been used to perform petrophysical inversion of surface wave data for global water content and temperature in the mantle transition zone (Meier et al., 2009), inversion of industrial seismic data sets for subsurface porosity and clay content (Shahraeeni and Curtis, 2011; Shahraeeni et al., 2012), inference of the Earth's 1D global average structure using body-wave travel-times (De Wit et al., 2013) and for earthquake detection and source parameter estimation (Käufl et al., 2014, 2015).

To produce a 3D shear-wave velocity versus depth model on any scale using the 2-step method, the inverse problem for structures with depth must be solved at many geographical locations (usually many thousands) over the area of interest. McMC inversion methods are computationally expensive and it is generally impractical to apply them in cases where parameters or data sets are large, where computational efficiency and processing time are usually limiting constraints due to the need to forward model many samples (of the order of thousands or millions) at each location. On the other hand, once trained, NNs and MDNs can often solve such inverse problems in seconds with no additional sampling. In addition, in cases where we wish to monitor changes in the subsurface, the same network can be applied rapidly to repeated data measurements, enabling the possibility of near-real time monitoring provided that the inputs to the networks can be

produced rapidly from the raw measured data. The aim herein is to investigate whether this is possible in practice.

In what follows I discuss the effect of data uncertainty and how to incorporate this within a neural network, then apply trained networks to field data from the Grane field in the Norwegian North Sea to create 2D shear-wave velocity maps of specific depth intervals. I compare the results from the MDN to non-linearized McMC methods, and thus prove that MDN surface wave inversion methods are both efficient and robust at the scale of reservoirs.

3.2 Method

3.2.1 Grane data

Grane is an offshore oil field in the Norwegian North Sea. A permanent reservoir monitoring (PRM) system was installed in 2014 over approximately 50 km^2 of the Grane seabed (Thompson et al., 2015). Ambient seismic noise is recorded continuously at the field using four-component sensors – 3-component geophones (Vertical, North and East) and a hydrophone. The data used in this study was preprocessed according to the protocol of Zhang et al. (2019), summarised as follows. Data from the vertical and hydrophone components were selected over a 6.5 hour interval. The data were bandpass-filtered between 0.35-1.5Hz and data from every pair of stations are cross-correlated using overlapping half-hour recording sections, then correlations are stacked over the full 6.5 hour interval. Cross-correlations of hydrophone and vertical component noise mainly contain information about Rayleigh-type waves. Phase velocities were automatically picked for the cross-correlation of each station-pair. Seventeen phase velocity maps and their corresponding standard deviation (uncertainty) maps were produced using eikonal tomography for periods between 0.6 to 2.2 seconds at 0.1 second intervals over a 50m x 50m grid. Figure 3.1a shows 4 examples of the phase velocity maps at periods 0.7s, 1.0s, 1.3s and 1.9s and their corresponding uncertainties.

Zhang et al. (2019) perform 1D, 2D and 3D Markov Chain Monte Carlo (McMC) Tomography over the Grane field to produce maps of the shear velocity

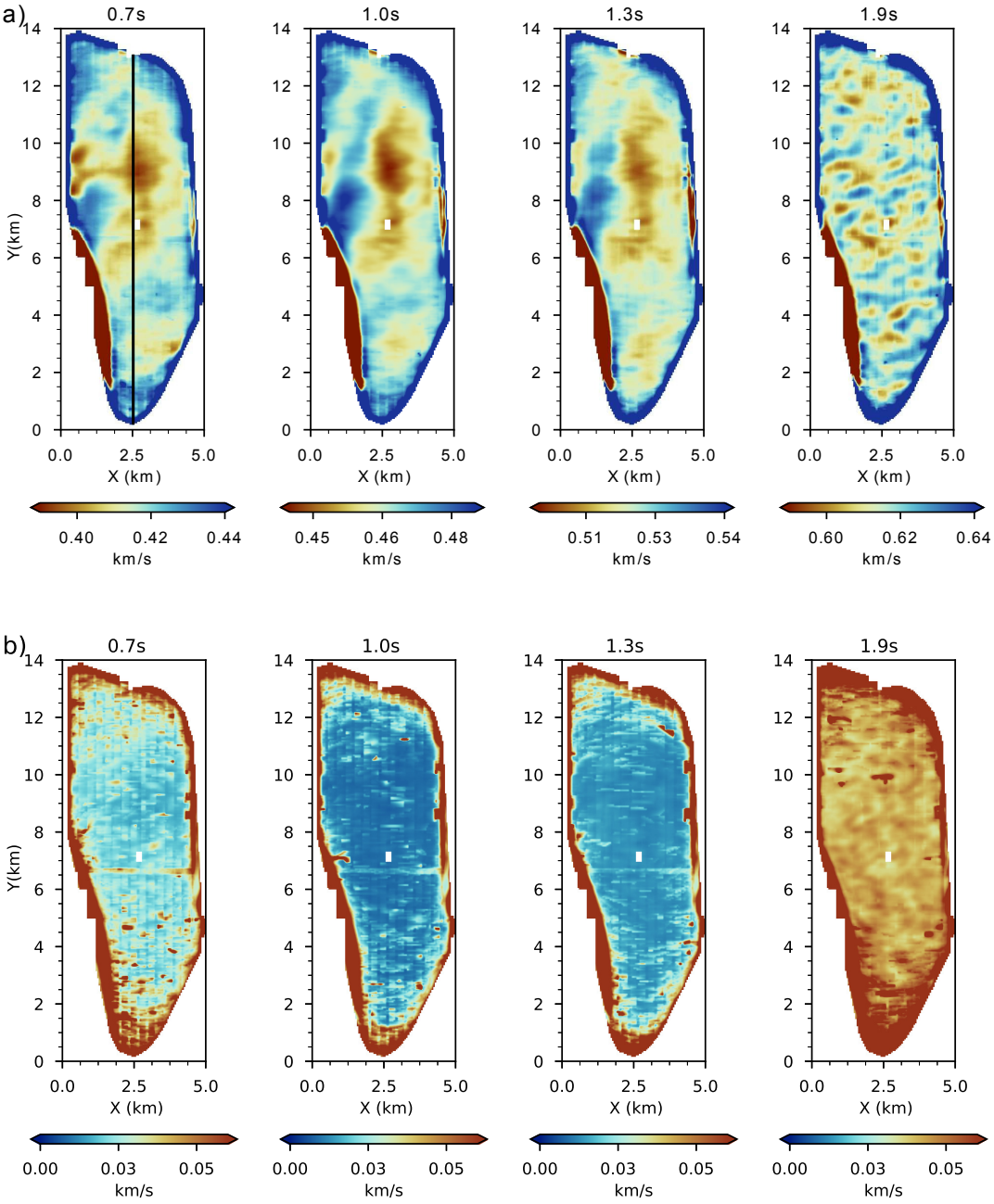

Figure 3.1 (a) A selection of four phase velocity maps used to compute discretised dispersion curves. Periods shown are 0.7s, 1.0s, 1.3s and 1.9s. (b) Maps of estimated standard deviation of uncertainties in the phase velocity at each location. Velocities and uncertainty colour scales are saturated at either end to prevent domination of outliers, and to highlight structure across the field. The vertical black line in the top left plot shows the location of a cross section shown in other Figures.

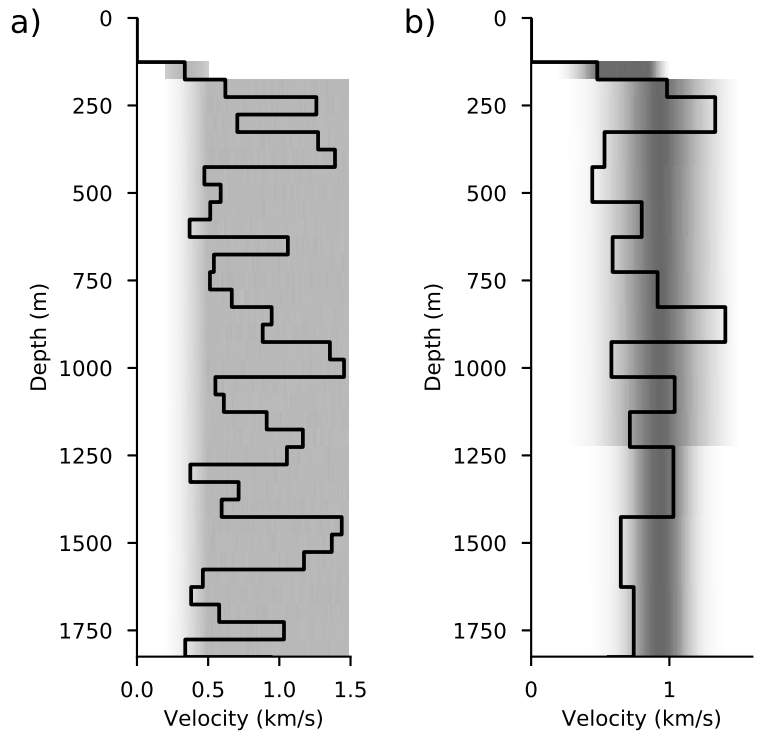

Figure 3.2 (a) Initial distribution of velocity structures created with a piecewise-constant discretisation over depth. (b) Distribution of velocity structures created after averaging structures in (a) over larger depth intervals. Grey-scale shows the probability density distribution, darker colours represent higher density of velocity structures, and the black line is an example of a randomly selected velocity structure in each panel which also illustrates the depth intervals used in cases (a) and (b).

structure with depth. However, McMC solutions are relatively slow to compute as they require ~10^6 3D or ~10^9 1D forward modelling simulations to obtain robust results and this needs to be repeated for each new data set for 4D surveys. Given that a PRM system exists, and that large amounts of data can be recorded in this area at many different points in time, a faster method is desired for monitoring applications.

3.2.2 Creating a Training set

The velocity structures **m** are parametrised as follows: each 1D structure has a water layer of 126m, followed by constant velocity layers every 25m to a depth of

100m below the water layer, then 50m thick layers down to 2000m below the water layer, beneath which there is a homogeneous half-space. For each velocity structure the S-wave velocity of the top solid layer was selected randomly from the Uniform probability distribution $v_{top} \sim U(0.2km/s, 0.5km/s)$ to represent unconsolidated near-surface sediments. For a fundamental mode surface wave to be observed, the top solid layer must have the lowest velocity (Galetti et al., 2017), therefore the following layers were randomly selected from distribution $U(v_{top}, 1.5km/s)$. I generated 1,000,000 velocity structures and Figure 3.2a shows the resulting distribution of velocities along with an example velocity structure. Selecting each velocity independently means that there are no correlations between each depth layer in the models. This keeps the prior broad, however if more geological information is known about the area then more sophisticated geostatistical methods can be used to create the training data. Spatial correlation between layers can be inferred from geological data such as well logs and then a set of velocity model examples can be created using a sampling technique such as Mote Carlo based methods.

The forward problem is solved for each of the generated velocity structures using the DISPER80 subroutines by Saito (1988) to obtain corresponding fundamental mode Rayleigh wave dispersion curves. The DISPER80 subroutines are fortran codes that calculate Rayleigh or Love wave group or phase velocities from a 1-dimensional layered model given the compressional velocity, shear velocity and density of each layer in the model. The phase velocities were calculated for periods 0.6-2.2s at 0.1s intervals in order to match the range available from ambient noise recorded at Grane. The DISPER80 forward modeller needs P and S-wave velocity and density for each layer in depth in order to calculate the phase velocities at each of any set of discrete periods. From the S-wave velocity structures calculated previously I compute a corresponding P-wave velocity v_p and the density ρ for each velocity layer based on typical values for sedimentary rocks using (Castagna et al., 1985; Brocher, 2005)

$$v_p = 1.16v_s + 1.36 \tag{3.1}$$

$$\rho = 1.74v_p^{0.25} \tag{3.2}$$

Rather than attempt to invert surface wave speeds at 17 periods for shear velocities in 40 depth layers, before training the velocity model is averaged over seventeen larger fixed-depth intervals (Figure 3.2b). Networks are then trained to invert for the velocity in each of these larger depth intervals.

3.2.3 Uncertainties

In past work, uncertainty information about the data is only included by adding random Gaussian noise to the training data set (Devilee et al., 1999; Meier et al., 2007a; Shahraeeni and Curtis, 2011; De Wit et al., 2013). Adding noise acts to regularise the network, helps to generalise when the network is inverting new data, and accounts for the data uncertainty in the Bayesian solution. However the disadvantage of such an approach is that when presenting the network with new data, updated uncertainty information for those particular data is not included in the inversion; indeed, that network would invert the data assuming that the incorrect (old) data uncertainties still pertain.

By contrast, here the data uncertainty is included as an additional set of inputs to the network. This makes sense because uncertainty is in fact additional pertinent information for each inversion. To train the MDN the clean synthetically-modelled data set is augmented with varying levels of noisy data. For each data point in the original synthetic data set a random percentage of noise ϵ is selected between the bounds outlined in Table 3.1 for six different Uniform distributions of ϵ. The noise is then added to the data according to

$$u_j = \epsilon \times d_j \tag{3.3}$$

$$\tilde{d}_j = \mathcal{N}(0,1)u_j + d_j \tag{3.4}$$

where u_j is the uncertainty value of the noisy data $\tilde{d}_j$ and $\mathcal{N}(0,1)$ is a Standard Normal distribution with mean 0 and standard deviation 1. An example of noisy data and the randomly chosen noise level is shown in Figure 3.3. Thus an augmented training set of data-velocity structure pairs $T_{uncer} = \{([\tilde{\mathbf{d}}_j, \mathbf{u}_j], \mathbf{m}_j) : j = 1, ..., N\}$ is generated, where the data consists of the noisy dispersion curves $\tilde{\mathbf{d}}_j$ and their associated uncertainties $\mathbf{u}_j$. The final data sets T and T_{uncer} are then

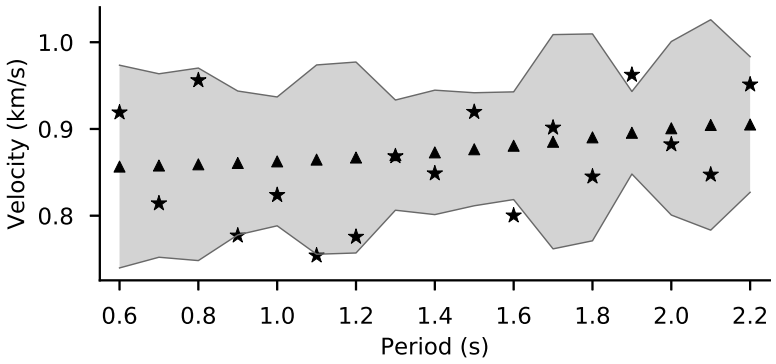

Figure 3.3 Graph showing a synthetic dispersion curve $\mathbf{d}$ (triangles) compared to a dispersion curve with added noise $\tilde{\mathbf{d}}$ (stars). The grey shaded area is the uncertainty $\mathbf{u}$ from Equation 3.3.

Noise Scenario	Percentage Noise % Min	Max
1	0	0
2	0	5
3	4	14
4	10	15
5	3	10
6	0	15

Table 3.1 Table of percentage range of noise added to data set in each of six different noise scenarios. Uncertainty is added to each datum according to Equations 3.3 and 3.4. In the first scenario uncertainties are zero.

shuffled and split into a training set (90% of training pairs) that is used to train the network for the optimum mapping, a validation set (5%) used during training to check the network is not over-fitting the training examples (see below), and a test set (5%) which is used post training to assess the network performance on previously unseen data. This final assessment provides weights E_i for the network ensemble in Equation 2.16. Early stopping is employed to prevent over-fitting of the network to the data: this is where the cost function is periodically checked on the validation set during training. When the cost function stops decreasing it is assumed that the network is already fit to the training data but is no longer improving its generalisation to new data. Training is then stopped.

3.3 Results

3.3.1 Network Design

Networks are trained for two different datasets: first a training set T in which data were perturbed by 10% Gaussian noise is used to train what is referred to herein as a *Noisy-MDN*. This MDN does not include uncertainty in its input vector. A second training set T_{uncer} includes a variable uncertainty vector as described in the previous section, which is used to train what is referred to as an *Uncertainty-MDN*. To include both the dispersion curve and their uncertainties in the latter networks two inputs are included as shown in Figure 3.4. The dispersion curve is passed through 2 layers, whilst the uncertainty is input separately and passed through one layer. The outputs of these two layers is then concatenated before being passed through two more layers to output the parameter vector that defines the probability distribution of the shear wave velocity structure.

Separate MDNs are trained for each depth interval in the velocity structure defined in Figure 3.2b. For each interval approximately 40 networks are trained from which the 10 networks with the lowest cost value across the validation set are selected for the ensemble . The weights and biases are randomly initialized using the Glorot uniform initializer (Glorot and Bengio, 2010) for each training run, and different sizes of layers in the different networks are used to create diversity. The different layer sizes were determined using a form of Bayesian optimization using the Python library hyperopt (Bergstra et al., 2015). Appendix A outlines the network configurations trained. The networks each use a Gaussian mixture with 15 kernels, so by using an ensemble of 10 networks a total of 150 kernels potentially contribute to each posterior distribution. However, I found that normally only 3 or 4 kernels with different means and standard deviations were assigned significantly non-zero amplitudes by each individual network.

3.3.2 Network Evaluation

A set of 100,000 synthetic velocity structures to which no network has previously been exposed were then created. These simulate relatively smooth velocity struc-

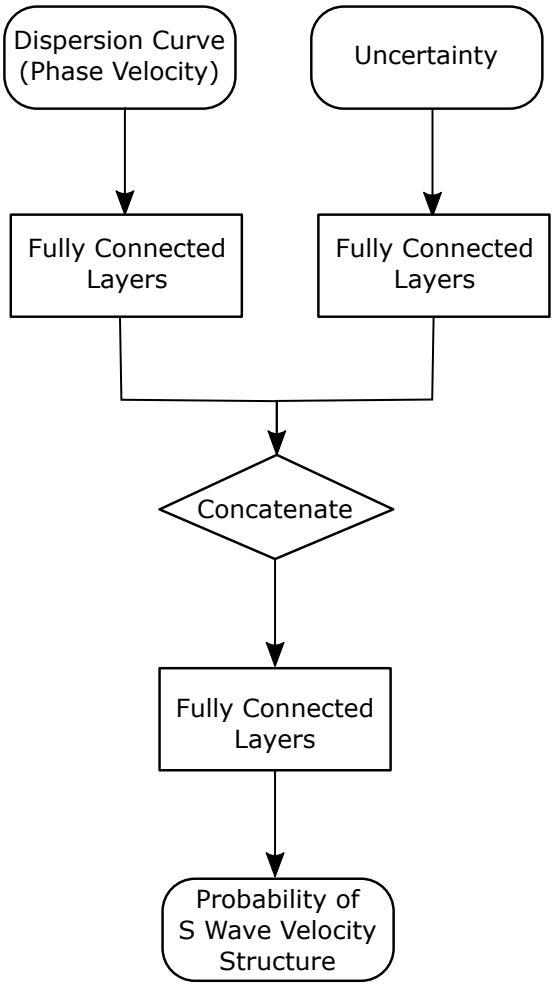

Figure 3.4 Diagram of network used to include uncertainty estimates in the input vector. Rounded edged boxes represent inputs/outputs of network. Squared edged boxes represent one or more fully connected layers within the model where the internal model weights are optimised during training. The structure of these is described in Table A.2. The diamond box represents the concatenation of layers: this step involves no new weights and simply concatenates the outputs of the previous layers. The arrows represent the direction of flow of data through the network.

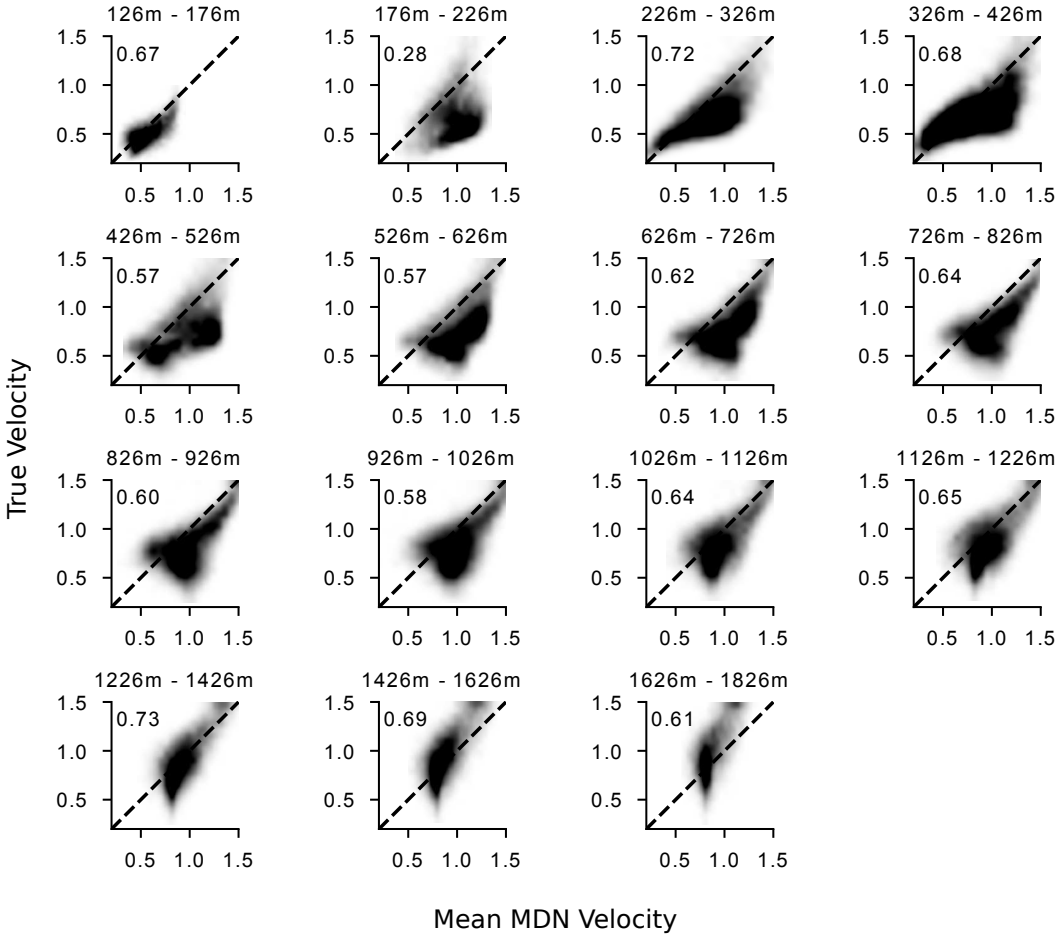

Figure 3.5 Mean of the posterior marginal pdfs from Noisy-MDN inversions, versus the true value of velocity for each velocity structure in the set of smooth models. Each graph represents a different depth interval as indicated above the graph. The corresponding Pearson correlation coefficient R is given in the top left corner of each graph.

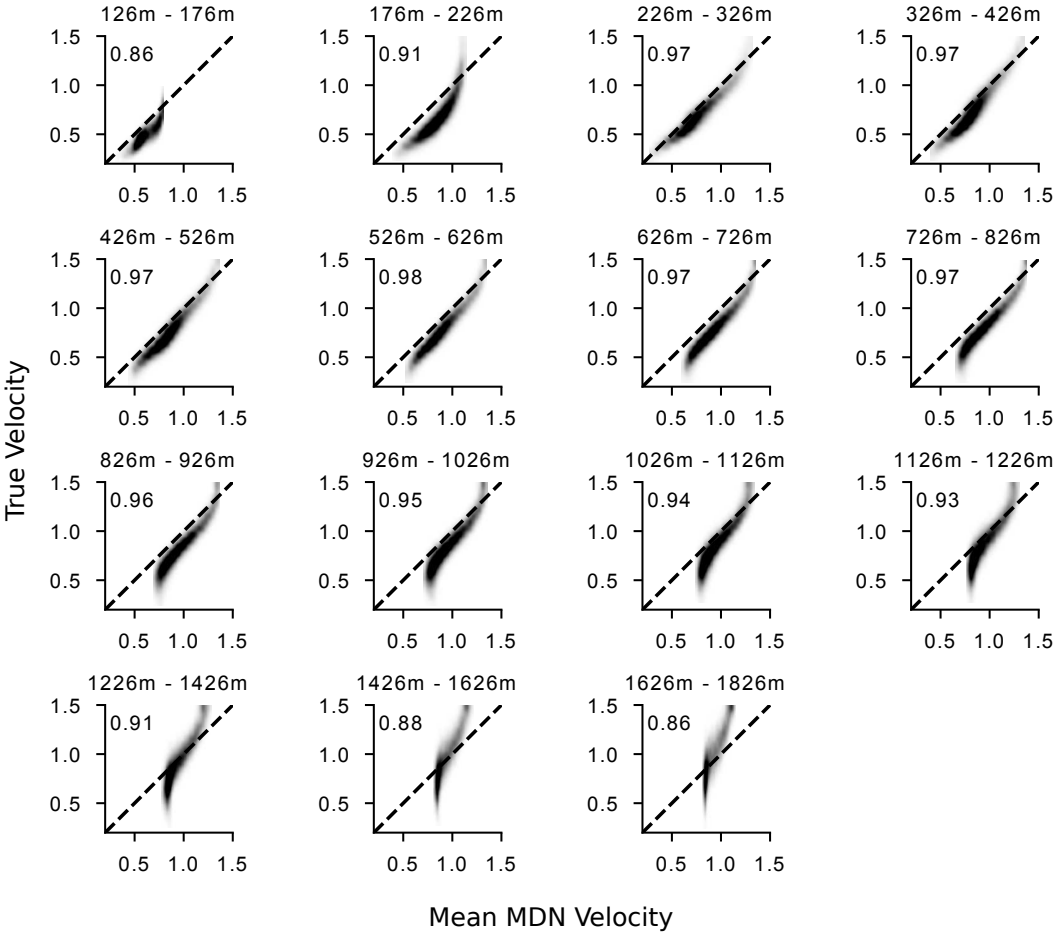

Figure 3.6 Mean of the posterior marginal pdfs from Uncertainty-MDN inversions, versus the true value of velocity for each velocity structure in the set of smooth models. Each graph represents a different depth interval as indicated above the graph. The corresponding Pearson correlation coefficient R is given in the top left corner of each graph.

tures by not allowing the velocity to vary more than 400m/s between neighbouring depth intervals. Corresponding data are created using the DISPER80 forward modeller, to which 10% Gaussian noise was added. For each depth interval in the velocity structure the MDN ensemble is applied to each of the 100,000 synthetic data and the mean of each posterior marginal pdf $\bar{\mu}_{post}$ is calculated by

$$\bar{\mu}_{post} = \sum_{i=1}^{M} \alpha_i \mu_i \tag{3.5}$$

The correlation between the mean of the posterior and the true target value for each data vector can be used to evaluate the performance of the networks when presented with new data. This evaluation does not use all of the information contained in each posterior pdf, but does provide a practical way to begin to evaluate network performance. Figure 3.5 shows the means of the posterior pdf of the fundamental mode Rayleigh wave Noisy-MDN inversions versus the true velocity values across all of the synthetic smooth velocity models, for each depth interval. The corresponding Pearson correlation coefficient, R, is shown in the top-left corner of the plot. The plots show a clear tendency for the mean of the network to over-estimate the true velocity value. When the same inversions are performed using the Uncertainty-MDNs (Figure 3.6) the correlation between the mean MDN velocities and the true velocities improves at every depth level. The additional information provided to the network that describes uncertainties in the data results in a significantly more accurate estimate of the velocity structure.

The plots in Figure 3.6 allow us to compare how the networks perform at different depth levels. The performance of the networks decrease with depth, and at the deeper levels (1626-1826m) the mean of the Uncertainty-MDN tends towards the mean of the prior. Figure 3.7 shows an example of the marginal posterior probability density function for two synthetic velocity structures at depths below 1226m. In both plots the true velocity structure is far away from the mean of the prior distribution yet the predicted marginal posterior distribution is very close to the prior: this shows that at these depths the networks are unable to add any information to the prior pdf given the data presented to the network. For this reason the following results are only shown down to a depth of 1226m.

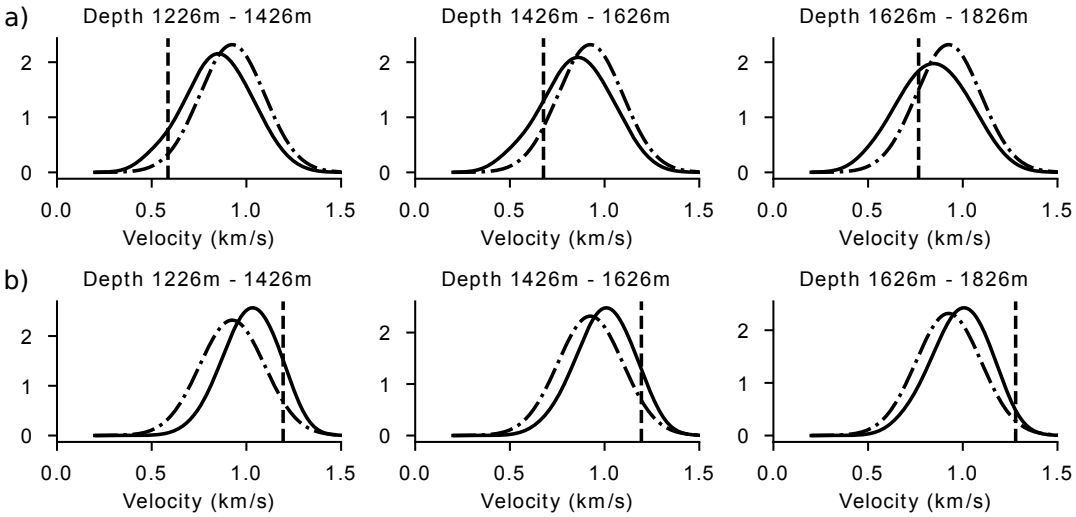

Figure 3.7 Individual probability density functions for depths below 1226m for two synthetic velocity structures in (a) and (b) respectively. The solid line is the marginal posterior probability density function from the MDN, the vertical dashed line is the true velocity value, and the dot-dash line is the prior probability density function.

3.3.3 Synthetic Results

Figure 3.8 shows the inversion of synthetic data from Noisy-MDN inversions: as seen in Figure 3.5 the networks generally over-estimate the predicted velocity and often the uncertainty does not encompass the true solution. The addition of noise of fixed standard deviation to training data examples does not fully represent the true uncertainty of each individual data point, so the inversion of noisy data results in unreliable estimates of velocity. However, when adding the uncertainty estimates explicitly into the network, Uncertainty-MDN results produce more reliable estimates as shown by the 1D depth inversions in Figure 3.9. The results from the networks are more representative of the true velocity structure, the maximum likelihood from the pdfs are much closer to the true shear wave velocity value, and the uncertainty ranges encompass the true velocity structure in all cases. This result is important for training and applying networks to field data. Noise, measurement error or assumptions about the Earth velocity structure can all affect the reliability of field data measurements; we never know whether our data is reliable, but we can often estimate the uncertainty in recorded data. Including

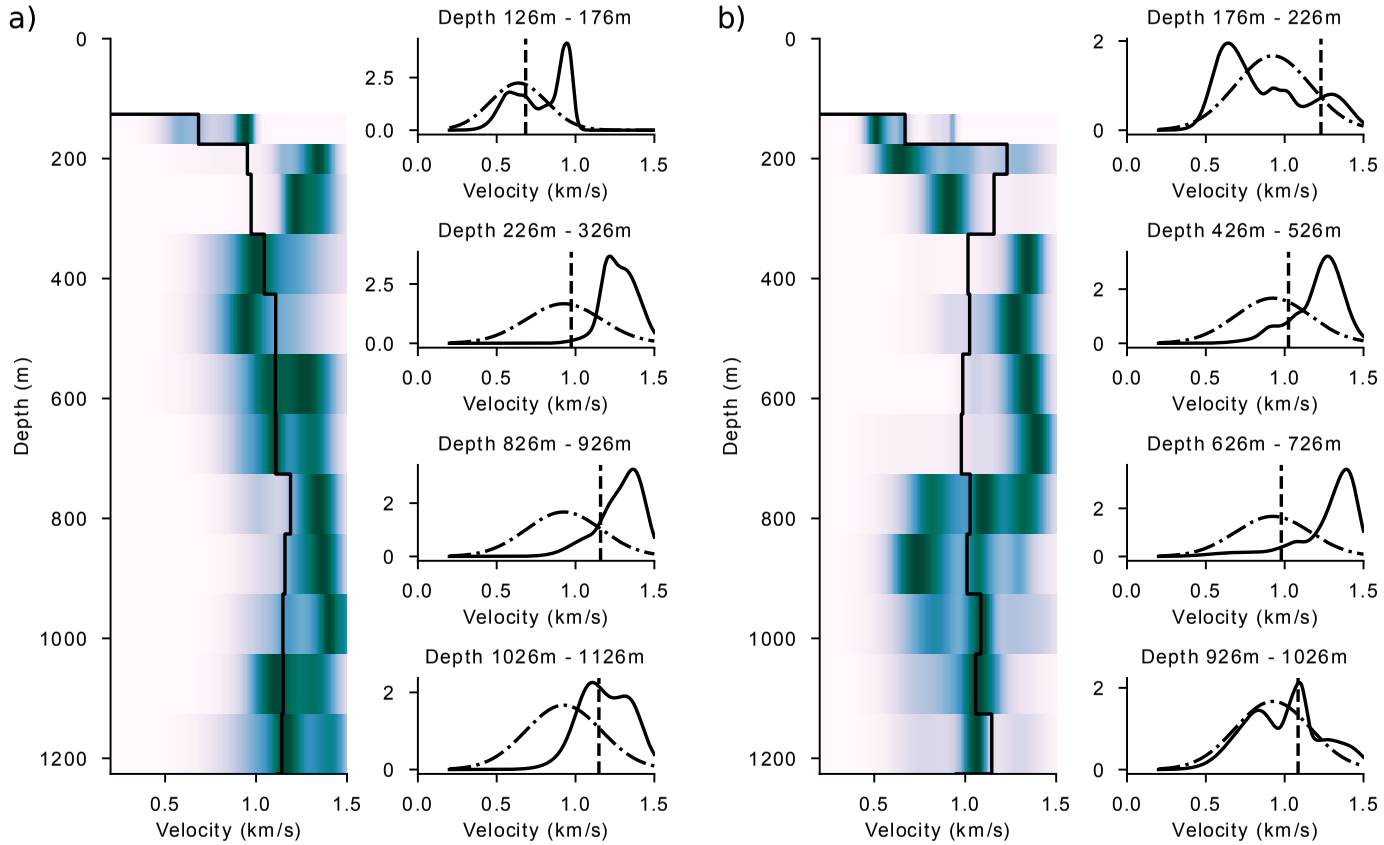

Figure 3.8 1D depth inversion result from Noisy-MDNs for two synthetic velocity structures with individual probability density functions shown for four depth levels. In the depth inversions dark colours represent areas of higher probability, each row of the posterior integrates to unity, and the black solid line is the true synthetic velocity structure. In the individual probability density functions the solid line is the marginal posterior probability density function from the MDN, the vertical dashed line is the true velocity structure, and the dot-dash line is the prior probability density function.

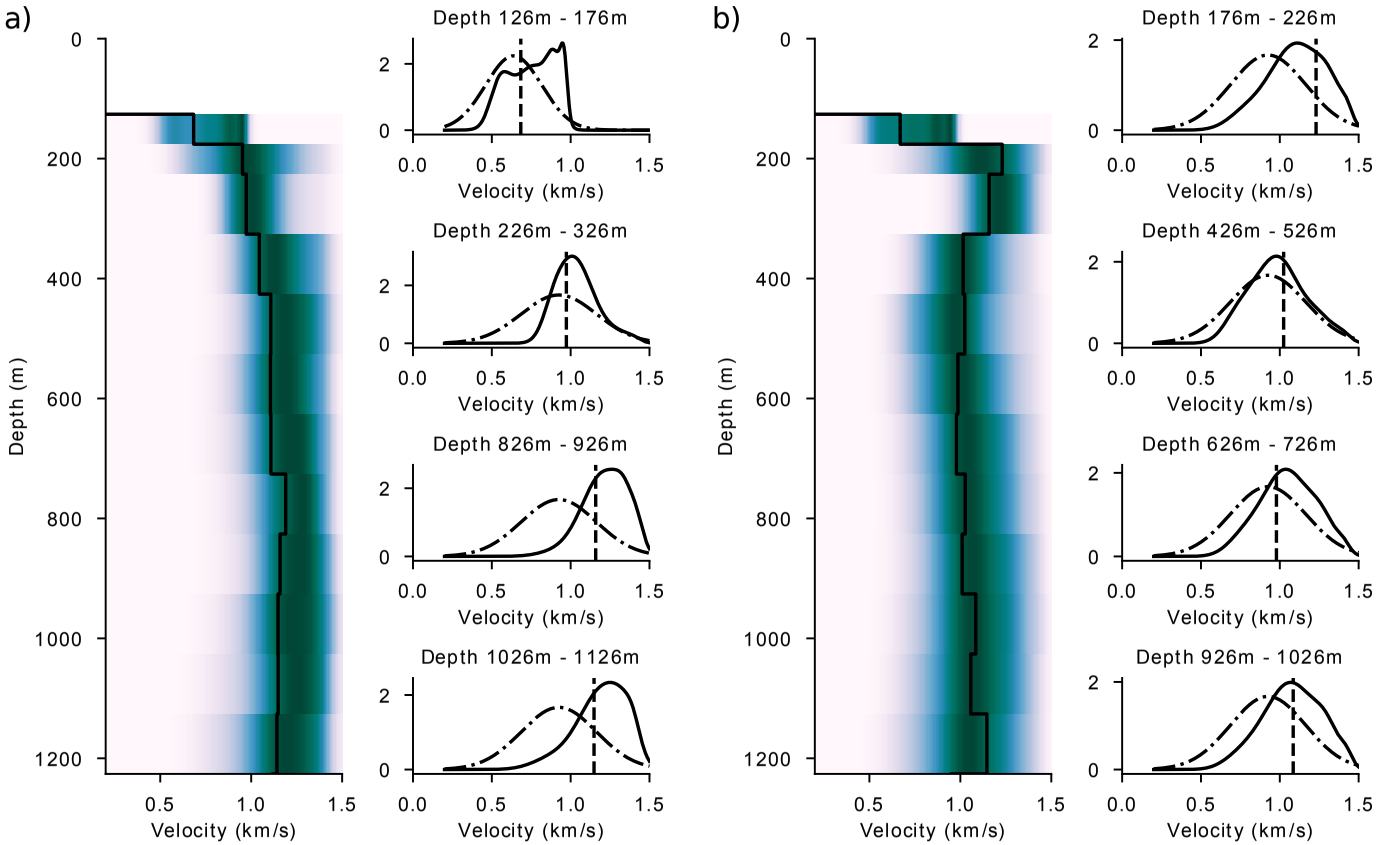

Figure 3.9 1D depth inversion result from Uncertainty-MDNs for two synthetic velocity structures with individual probability density functions shown for four depth levels. In the depth inversions dark colours represent areas of higher probability, each row of the posterior integrates to unity, and the black solid line is the true synthetic velocity structure. In the individual probability density functions the solid line is the marginal posterior probability density function from the MDN, the vertical dashed line is the true velocity structure, and the dot-dash line is the prior probability density function.

the uncertainty information in networks allows them to be applied to field data with increased confidence that they will produce reliable estimates of posterior pdfs that encompass the true subsurface velocities.

3.3.4 Field Data

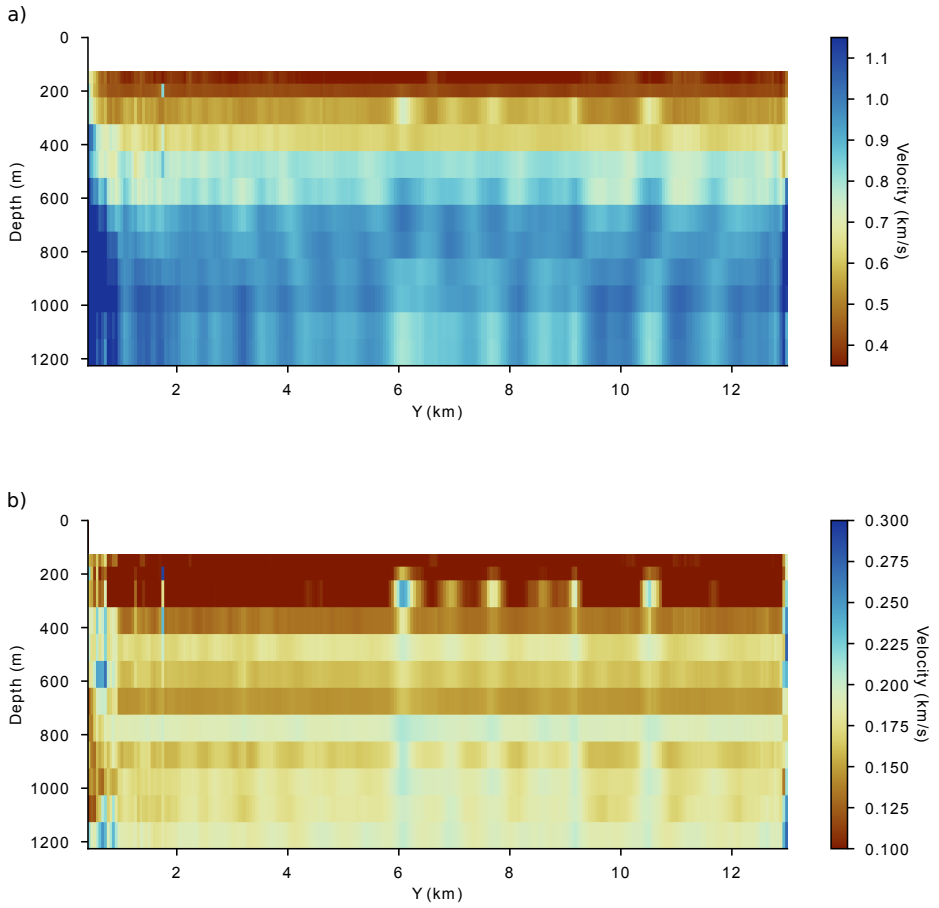

Figure 3.10 (a) Mean shear velocity cross-section, and (b) corresponding posterior standard deviation cross-sections from the Noisy-MDN inversion. The top white layer represents the water layer where shear velocity is zero.

The final trained MDNs are applied to invert Rayleigh wave phase velocities from the Grane field in the Norwegian North Sea. Dispersion curves were extracted at each grid point producing 26,772 dispersion curves to be inverted for 1D depth-velocity structures. The standard deviations shown in Figure 3.1b were extracted at each point and used as the uncertainty vector input to the

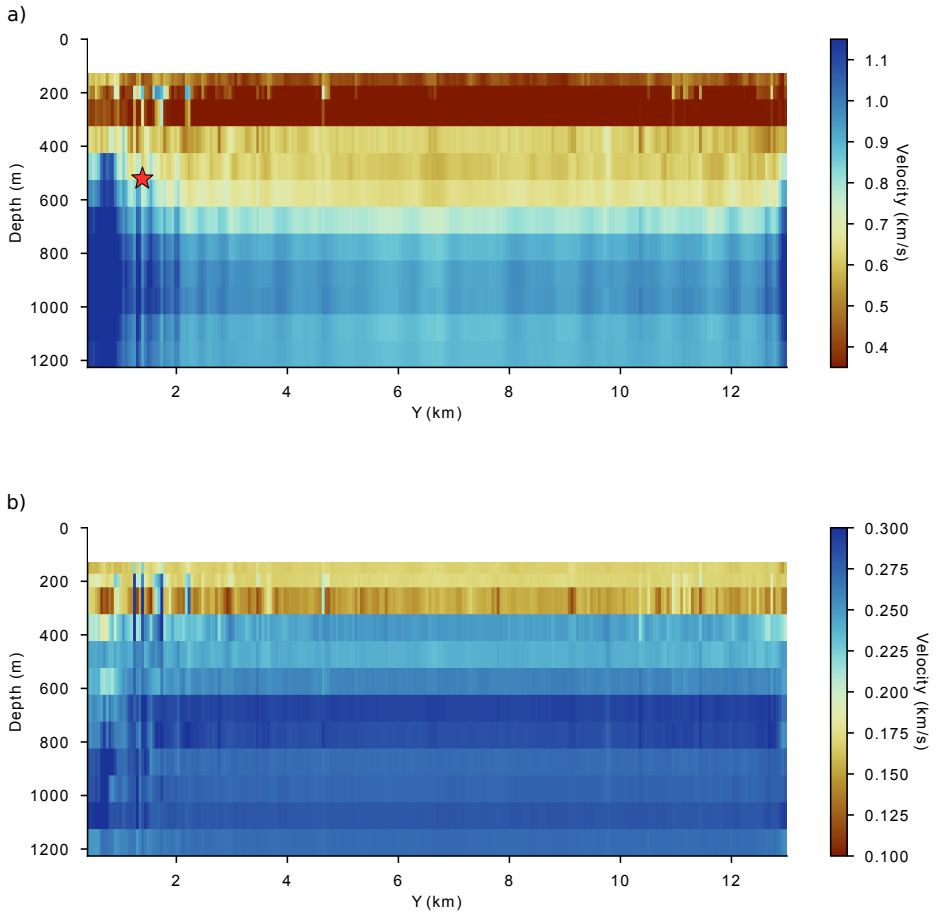

Figure 3.11 (a) Mean shear velocity cross-section, and (b) corresponding posterior standard deviation cross-sections from the Uncertainty-MDN inversion. The top white layer represents the water layer where shear velocity is zero. Red star shows location of the joint pdf shown in Figure 5.7.

Uncertainty MDNs (Figure 3.4). Figures 3.10 and 3.11 show the mean and associated standard deviations (representing uncertainty) of the posterior pdf estimated at the location of the black line in Figure 3.1a from Noisy-MDNs and Uncertainty-MDNs respectively. Both plots of the mean show a reasonably similar structure: a near-surface low velocity layer down to 300m, then an increased velocity down to 600m, with yet higher velocities below this. However, the layers are more distinct in the inversion using the Uncertainty-MDNs. Figure 3.10a from the Noisy-MDN shows a higher variability in the velocity below 600m than does the mean in Figure 3.11a, and the velocity highs in Figure 3.10a coincide with

higher uncertainties in Figure 3.10b. When networks are trained including the full uncertainty information (Figure 3.11) these velocity highs disappear so that the mean velocity and uncertainties are laterally smoother across the section. I therefore now focus on the Uncertainty MDN results.

Figure 3.12 shows the mean and standard deviation horizontal depth slices from the Uncertainty-MDNs. In the near surface maps (126m - 326m) the results show similar structures to those in the phase velocity maps in Figure 3.1a at short periods, for examples within the dotted red boxes in Figure 3.12a. The deeper maps (536m - 826m) show structures similar to that of the longer period phase velocity maps, but also a higher standard deviation (Figure 3.12b) than shallower layers. As a result, the shear velocity variation in these deeper structures falls within their standard deviation, suggesting that they might not represent true structure.

The method outlined above can easily be extended to joint inversion of fundamental and first higher mode data by adding two further inputs: the vector of first higher mode phase velocity values generated from the velocity structures in the original training set and a vector of their associated uncertainties. Figure 3.13 shows the cross-section results and Figure 3.14 shows the results from 4 depths layers, 126m-176m, 226m-326m, 426m-526m, 626m-726m, from Uncertainty-MDN joint inversion. The same features seen in the shallow layer of Figure 3.12a are seen in the shallow layer of the joint inversion, highlighted by the red dashed boxes in Figure 3.14a. However, the velocities are on average higher than the fundamental mode-only MDN inversions and the standard deviations are larger. In addition, the depth slice at 226m-326m is entirely different to the corresponding slice in Figure 3.12a. Figure 3.13a shows that the low velocities observed in the top layers of the Uncertainity MDN fundamental model inversion (Figure 3.11a) no longer exist in the joint inversion with higher modes, showing the latter waves appear to have added additional information to the inversion. However, I am less confident about the quality of the higher mode dispersion measurements than those from the fundamental mode, so I include this result as a demonstration, but in the Discussion below I focus mainly on the fundamental mode results.

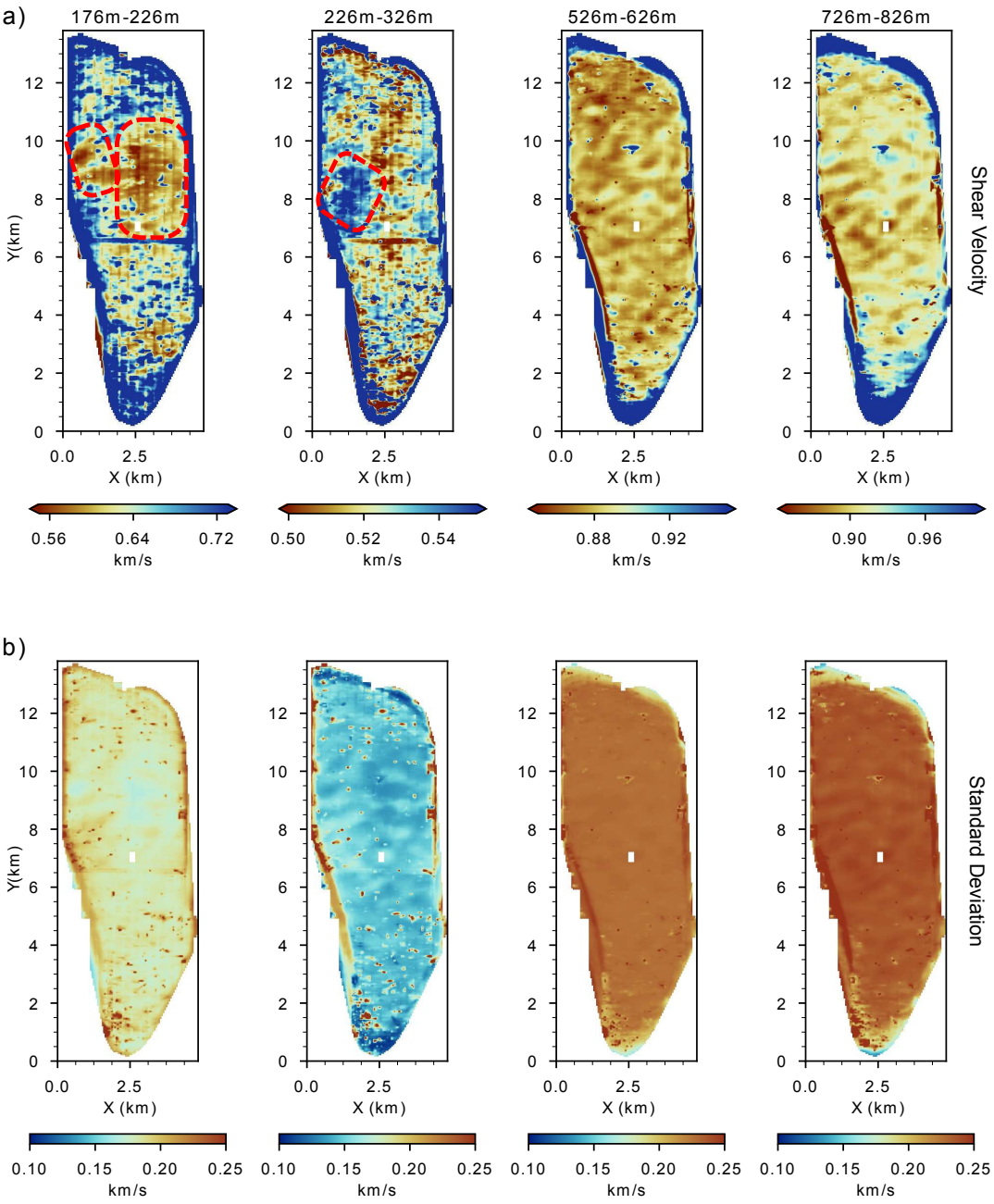

Figure 3.12 Fixed depth maps of (a) the mean and (b) the standard deviation of the shear velocity from Uncertainty-MDN inversion of fundamental mode Rayleigh dispersion at depth slices 176m-226m, 226m-326m, 526m-626m, 726m-826m.

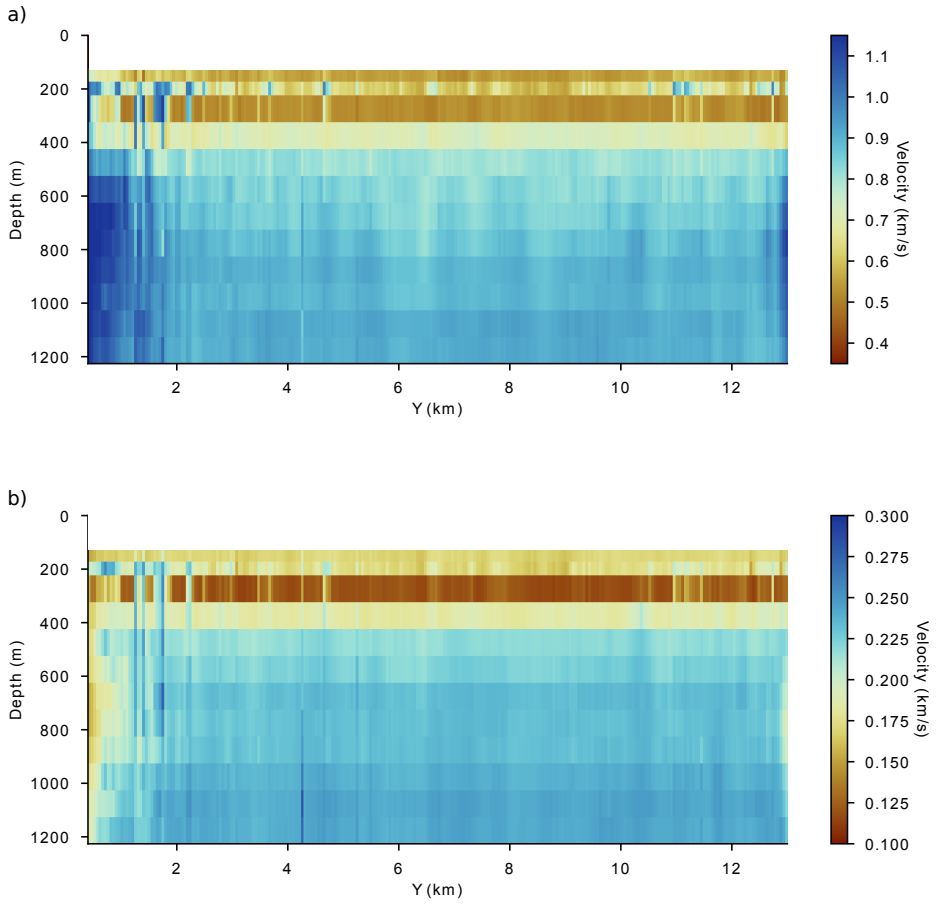

Figure 3.13 (a) Mean shear velocity cross-section, and (b) corresponding posterior standard deviation cross-sections from the Uncertainty-MDN inversion of fundamental and first higher mode Rayleigh dispersion. The top white layer represents the water layer where shear velocity is zero.

3.4 Discussion

I inverted Rayleigh wave phase dispersion curves for subsurface shear-wave velocities using MDN's trained with added Gaussian noise at a fixed standard deviation to simulate average data uncertainties, and a second type of MDN with the data uncertainty vector included as an additional input. I showed that to invert noisy data for reliable velocity structures the uncertainty estimates should be included in the network.

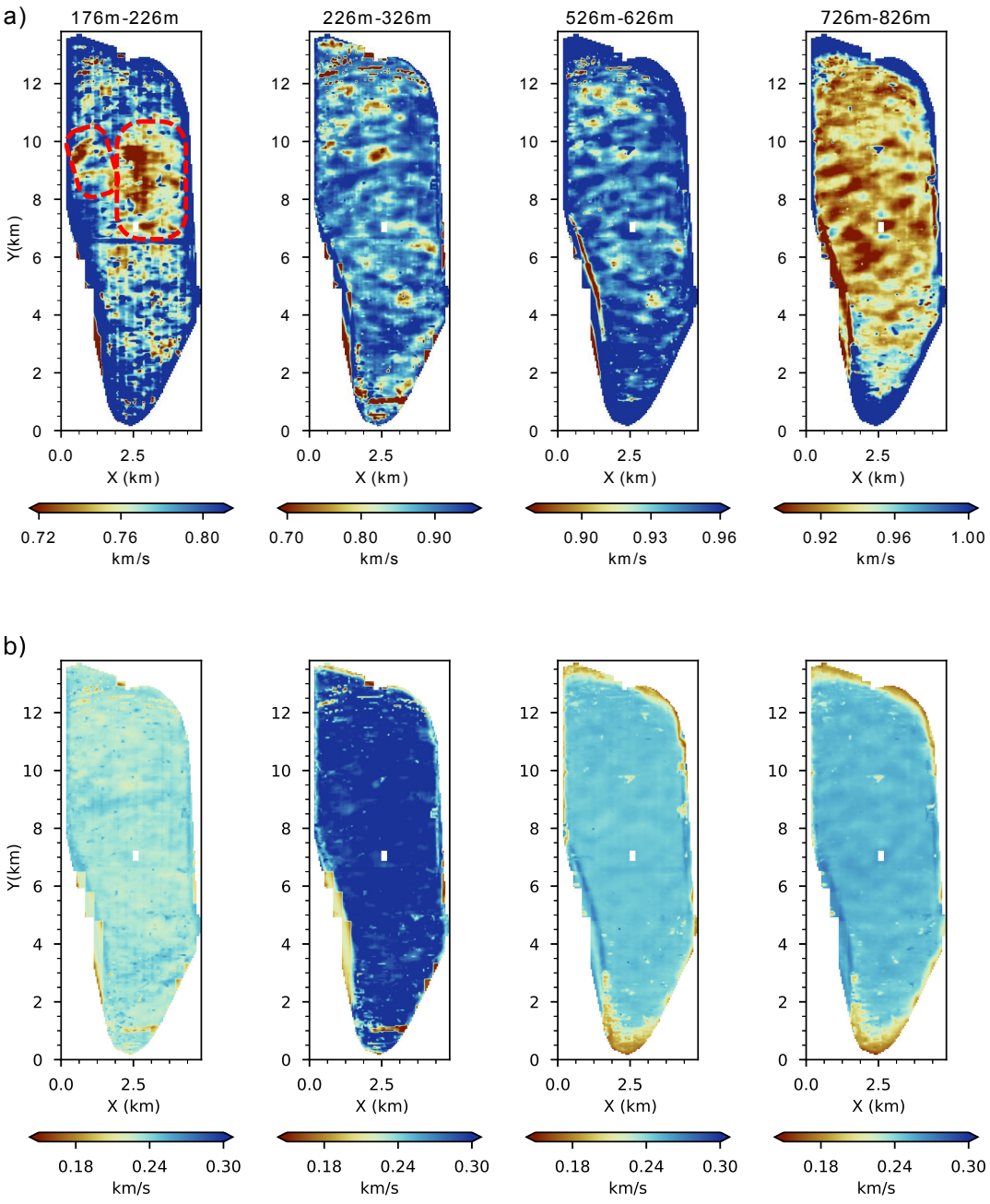

Figure 3.14 Fixed depth maps of (a) the mean and (b) the standard deviation of the shear velocity from Uncertainty-MDN inversion of fundamental and first higher mode Rayleigh dispersion at depth slices 176m-226m, 226m-326m, 526m-626m, 726m-826m.

	Noisy-MDN	Uncertainty-MDN
1D McMC	0.0018	0.0025
2D McMC	0.0025	0.0026
3D McMC	0.0021	0.0019

Table 3.2 Table showing the mean-squared difference between the Noise- and Uncertainty-MDN inversion cross sections, and the Markov Chain Monte Carlo cross sections of Zhang et al. (2019).

A constant number of fixed depth-velocity intervals were used in each MDN inversion, leading to inversions for effective medium (averaged) shear velocities for each fixed depth interval. A trans-dimensional network inversion would have had to include varying depths and number of layers which would significantly increase the dimensionality of the network inversion problem and require a much larger training set and more complex network structure. This in turn would increase training time and the memory needed for training, and would likely make the network outputs less stable and reliable since the posterior would effectively be emulating the inverse function in a higher dimensional space. For the intended application (to test the ability to rapidly monitor the overburden of a permanently instrumented field), the inversion for effective medium parameters over fixed depth intervals was sufficient.

3.4.1 Comparison with Monte Carlo Methods

I compare the Noise- and Uncertainty-MDN inversion results to the Markov chain Monte Carlo results of Zhang et al. (2019). Figure 3.15 shows the mean shear velocity cross sections of Figures 3.10a and Figures 3.11a along with the same cross sections from 1D, 2D and 3D trans-dimensional Markov chain Monte Carlo inversion (McMC). Despite comparing a trans-dimensional result from Monte Carlo methods with fixed-depth layer results from MDNs, all cross sections show a similar, approximately 3-layered structure. The 1D McMC (Figure 3.15c) most represents the networks trained using the Noisy-MDNs (Figure 3.15a) as both contain vertical velocity anomalies in the deeper part of the section. The Uncertainty-MDN has smoother variations laterally but also has a thicker near

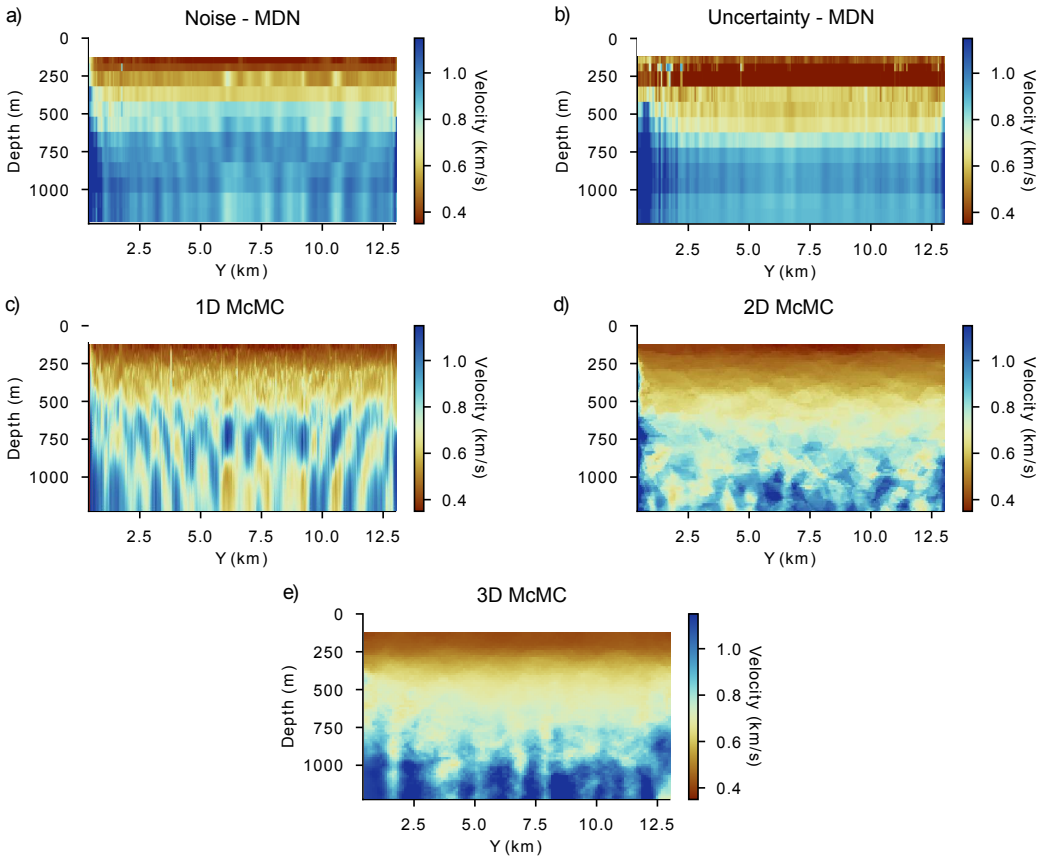

Figure 3.15 Mean shear velocity along the cross-section in Figure 3.1a from (a) MDN inversions using a training set with added Gaussian noise of fixed standard deviation, (b) MDN inversions using estimated data uncertainties as added input data, (c) independent 1D Monte Carlo inversions, (d) a single 2D Monte Carlo inversion, and (e) a 3D Monte Carlo inversion, where results in (c), (d) and (e) are from Zhang et al. (2019). The top white layer represents the water layer, where the shear velocity is zero.

surface velocity layer and the second layer extends deeper into the section (to ∼700m); this is more representative of the 2D and 3D McMC results (Figure 3.15d and 3.15e). This is confirmed by examining the mean-squared difference (MSD) between the mean of each MDN inversion and the Monte Carlo inversions in Table 3.2: the Noisy-MDN has a lower MSD compared to the 1D McMC inversion and the uncertainty MDN has a lower MSD compared to the 3D McMC inversion. This implies that by adding uncertainties to the MDN training we allow smoothness in the mean estimates which the 3D McMC results suggest is reasonable across Grane.

3.4.2 Joint Posterior Probability Density Functions

The results in the previous section are created from the 1D marginal posterior pdf $p(m^i \mid \mathbf{d})$ of the shear velocity in each layer independent of other velocities in each 1D profile. The correlations between velocities at different depths cannot be derived from such results. To estimate correlations it is necessary to analyse the joint posterior pdf $p(m^i, m^{i+1} \mid \mathbf{d})$, which can be constructed from the product of the conditional and marginal pdfs.

$$p(m^i, m^{i+1} \mid \mathbf{d}) = p(m^i \mid \mathbf{d}) \times p(m^{i+1} \mid m^i, \mathbf{d}) \tag{3.6}$$

The marginal pdfs $p(m^i \mid \mathbf{d})$ are given by the results shown in the previous sections. New networks are trained to estimate the conditional pdfs $p(m^{i+1} \mid m^i, \mathbf{d})$ by extending the input vector of the data with the velocity to which we want to condition the data: in this example this is the velocity of the layer above the one being estimated.

Figure 3.16 shows the results from the location shown by the red star in the Grane cross-section from Figure 3.11a. The plot shows a weak negative correlation, representing the weak trade-off between velocities in subsequent layers. This is likely to be because a relatively coarse parametrisation (compare that in Figure 3.2a and 3.2b) was used over depth for the inverse problem. If a finer parametrisation was used, such trade-offs would emerge more strongly as demonstrated by Meier et al. (2007b).

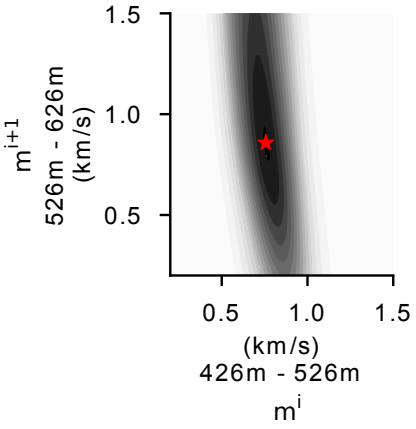

Figure 3.16 Joint pdf comparing the velocity trade-off between two adjacent layers m^i and m^{i+1} at depths given in the axis labels. The red star represents the mean velocity shown in Figure 3.11a.

3.4.3 Inversion Speed

Post-training, neural networks invert new data extremely rapidly: in this study it took approximately 21 CPU seconds to invert all 26,772 locations. The results are compared to Monte Carlo methods which are known to be computationally expensive (Bodin and Sambridge, 2009): the McMC methods used to create the crosslines shown in Figure 3.15 took approximately 186 CPU hours for 1D, 206 CPU hours for 2D, and 4824 CPU hours for 3D inversions. Despite the higher vertical resolution of results from McMC methods (since the parametrisation over depth varies in those inversions), the compute-time for inversions is between 4 and 6 orders of magnitude larger than for trained MDNs.

A comparison of time per inversion of an individual location for 1D MDN, 1D Monte Carlo and a 1D linearized inversion is shown in Figure 3.17a. Monte Carlo inversion is computationally the most expensive, and MDN inversion is two orders of magnitude faster than even linearized inversion. However, in this comparison I only accounted for the speed of the inversion which is not the full computational expense involved in using neural networks. Training a network before the inversion takes significant computational time: in this study it took 1280 CPU hours to train all of the networks used. It should be noted that training

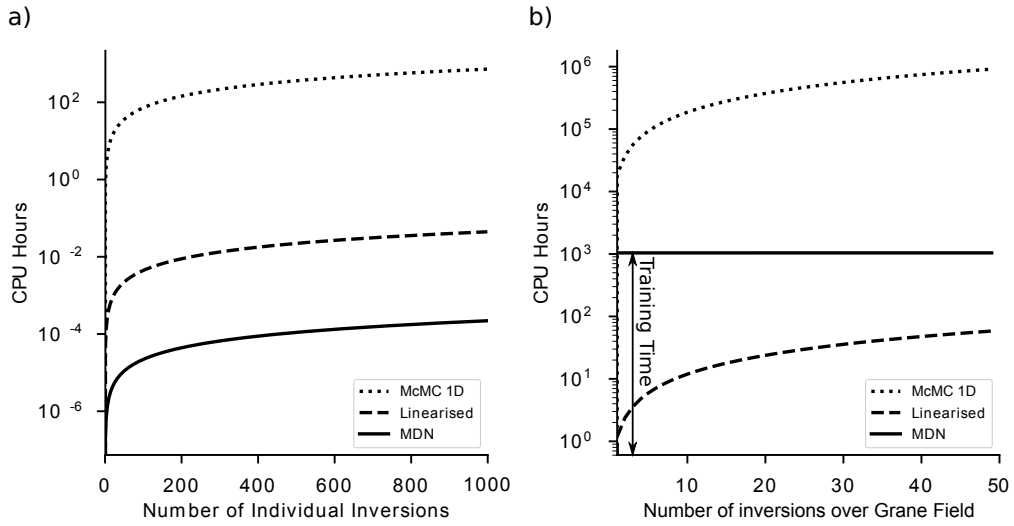

Figure 3.17 Plots showing the CPU hour time for (a) one inversion per location and (b) the inversion over the entire Grane field using MDNs, linearized inversion and Monte Carlo 1D (McMC 1D) methods. The typical time taken to train the MDN (including the forward modelling runs) is included in (b) but not in (a).

the network needs only to be done once, and hence is independent of number of locations to be inverted; therefore inverting more locations renders the MDN inversion method more computationally efficient. Figure 3.17b compares the CPU hours needed for monitoring-style repeated inversions across the full Grane field as performed in this study, including the time required for training MDNs. The initial cost of training a network before the first inversion is high, but thereafter repeated inversion of new data sets is nearly free. In comparison to 1D McMC methods, even accounting for the initial training period, neural network methods are more efficient. Linear inversion methods are computationally cheaper that MDN methods: in this case approximately 1000 inversions of the same field would be needed for the neural network method to become cheaper. Surface wave tomography is a non-linear inversion problem and despite the linearized inversions involving fewer CPU hours they can only give approximate solutions, in particular for uncertainty estimates, due to their implicit assumption of incorrect (linear) physics. The neural network method provides a fully probabilistic, fully non-linear solution that, once a network is trained, can be used to obtain rapid, repeated inversions.

3.5 Conclusion

I trained mixture density networks (MDN's) to invert fundamental mode Rayleigh wave dispersion curves for subsurface shear-wave velocity using two different methods to represent data uncertainties. The MDNs give a fully probabilistic solution to this non-linear inversion problem giving comparable results to Monte Carlo solutions. I show that inputting data uncertainties explicitly to the network provides a more reliable solution estimate on noisy synthetic data, and a smoother result that is more similar to 3D Monte Carlo inversion results on field data. The same method is used for joint inversion of fundamental and first higher mode data. Once trained, the neural network approach gives rapid results that can be repeatedly applied to similar types of data in monitoring scenarios.

Chapter 4

Near-real time near-surface 3D seismic velocity and uncertainty of ambient seismic noise

With the development of large and dense seismic arrays the nearly-direct measurement of local medium properties is possible by 3D or volumetric wave equation inversion, while 2D wave equation inversion provides direct measurements of surface wave dispersion. Computationally efficient imaging and inversion methods are needed to utilise the large amount of such information from these arrays. By using the second-order derivatives of the temporal and spatial wavefield, short recordings can be inverted using the Helmholtz wave equation to give frequency-dependent point phase velocity estimates. A class of neural networks called mixture density networks are trained to invert dispersion curves for 1D depth-velocity profiles and their uncertainties. Once trained, the networks are able to produce 3D depth-velocity profiles in a matter of seconds. The full inversion process from field data to 3D depth-velocity structure is computationally cheap, opening up the possibility of near-real time monitoring using dense arrays.

Author Contributions: This chapter is based on Cao et al. (In Press) of which I am second author. Sections 4.2.1, 4.3 and 4.4.1 are a summary of work done by the co-authors of the paper and their results. Work and results presented in Sections 4.2.2 and 4.4.2 are entirely my own and use the results of Section 4.4.1.

4.1 Introduction

Seismic interferometry has long been used to image surface waves by cross-correlating recordings from two stations (Aki, 1957; Claerbout, 1968; Curtis et al., 2006; Wapenaar and Fokkema, 2006). Applications range from imaging on a global (Shapiro and Campillo, 2004; Ruigrok et al., 2008), regional (Campillo and Paul, 2003; Gerstoft et al., 2006; Lin et al., 2008; Nicolson et al., 2012, 2014) and reservoir scale (Stewart, 2006; Bussat and Kugler, 2011; Allmark et al., 2018). The advent of large and dense arrays full recording of wavefields with a high spatial and temporal resolution mean that the nearly-direct measurement of local medium properties is possible without having to perform interferometry. This is possible by interpreting the gradients of the wavefield using a wave equation or its closed form solution a method often referred to as seismic gradiometry. Curtis and Robertsson (2002) first showed that one can estimate P and S wave velocities from derivatives of a wavefield using volumetric recordings (arrays distributed so as to span 3D space) which allows spatial wavefield recordings to be calculated horizontally or in depth. Since then, a number of methods have emerged to infer information from wavefield gradients.

Under the assumption that the wavefield is composed of non-overlapping plane waves Langston (2007a,c,b) use the inversion of first-order temporal and spatial derivatives, known as *wavefield gradiometry*, to determine wave attributes such as geometrical spreading, horizontal slowness and changes in radiation pattern. This latter approach was extended to 3 dimensions by Poppeliers et al. (2013) who showed that wavefield gradiometry is sensitive to uncorrelated noise and the presence of interferring waves. Gradiometry has been applied on a continental scale (Liang and Langston, 2009; Liu and Holt, 2015), at vertical boreholes (Langston and Ayele, 2016), and for terrestrial and lunar near-surface studies (Edme and Yuan, 2016; Sollberger et al., 2016).

Gradiometry is limited by the assumption that the plane wave arrivals are non-interferring. For more complex wavefields, the phase velocity can be estimated through the inversion of an eikonal equation (Eikonal tomography) using the spatial gradients of traveltimes of surface wave arrivals as data (Lin et al., 2009;

Liu and Holt, 2015). This method has been applied to cross correlations of ambient noise recorded over the Valhall field to produce images of surface wave velocity (de Ridder and Dellinger, 2011; Mordret et al., 2013b) and to include azimuthal anisotropy (Mordret et al., 2013a). Lin and Ritzwoller (2011) use earthquake surface wave data to image isotropic and anisotropic structures in the crust beneath western United States. These methods involve picking arrival times and so requires recordings from large earthquakes or cross-correlations of long seismic noise recordings, to be able to obtain accurate time picks.

By directly inverting the Helmholtz equation for phase velocity maps using spatial and temporal gradients of ambient seismic noise, de Ridder and Biondi (2015) avoided the need to cross-correlate long recordings for first arrival travel times. They showed that using only 10 minutes of recordings they could retrieve Scholte wave velocities with comparable results to that of seismic noise cross-correlation tomography. They and Muijs et al. (2003) found that the errors in second-order spatial derivatives needed for Helmholtz or wave equation inversion cause significant errors in the final velocity estimates and accurate finite difference stencils are necessary. Helmholtz equation inversion was extended to anisotropic media by de Ridder and Curtis (2017), and Zhan et al. (2018) showed that Helmholtz equation inversion can be improved by combining the method with compressive sensing; in that case this approach was shown to yield better results than Eikonal tomography.

Neural networks approximate a non-linear mapping between two parameter spaces. By presenting the network with a set of data-model pairs, it can be trained to create a mapping from the data to the model parameter space. This is particularly useful in geophysical inverse problems where the forward mapping, from model to data parameter space, is well known or simple to calculate (to construct training data) but when the inverse mapping is complex or difficult to determine directly. Once a network has been trained it can be given previously unseen data and will output a new model estimate in seconds. Using a specific class of neural network, called a mixture density network (MDN), the network can also provide uncertainties in the parameter estimates (Bishop, 1995). Meier et al. (2007b,a) have used MDNs to invert regional surface wave dispersion curves

to give fully probabilistic estimates of global crustal thickness models. MDNs have also been used to perform petrophysical inversion of seismic data sets for subsurface porosity and clay content.

In what follows I outline the gradiometry theory used in the chapter. I present the phase velocity maps from gradiometry and mean and standard deviation maps at specific depth levels from the neural network inversion. I compare the velocity-depth results to Monte Carlo sampling methods and show that neural network inversion results give comparable results but at a much smaller computational cost.

4.2 Method

4.2.1 Wavefield Gradiometry Method

If wavefield recordings are dominated by far-field fundamental mode surface waves then the wave propagation is approximately described by the two dimensional dispersive scalar wave equation

$$\nabla^2 \hat{u}(x,y,\omega) + \hat{s}^2(x,y,\omega)\omega^2 \hat{u}(x,y,\omega) = -\hat{f}(x,y,\omega) \tag{4.1}$$

where u is the scalar wave field varying in time and space, f is a generalised source term, ∇^2 is the Laplace operator acting on the two spatial dimensions, s is the slowness and a hat symbol above a quantity represents the frequency domain Fourier transform of that quantity. Following the method of de Ridder and Biondi (2015) the recordings are band-passed using a narrow frequency-domain Hann window with a central frequency ω' and thus we can ignore the frequency dependence of the phase velocity so $\hat{c}(\omega) = c_{\omega'}$. It is assumed that no strong local sources are acting within the area of recording since the source distribution is generally unknown for ambient seismic noise so that $\hat{f}(x,y,\omega) = 0$. The filtered recording would then obey the time domain equation:

$$\nabla^2 u_{\omega'}(x,y,t) = \partial_t^2 u_{\omega'}(x,y,t) s^2_{\omega'}(x,y) \tag{4.2}$$

where ∂_t^2 is the second-order derivative acting in the time domain. This equation relates the temporal and spatial derivatives of the wavefield to the phase velocity, for a narrow-bandwidth filtered wavefield.

It is not usually possible to observe the second order gradients of the wavefield directly. However, they can be estimated using finite difference approximations to the derivative operators. Assuming the data is recorded on a regular grid with interval sizes Δx and Δy in the x and y directions and the temporal data sampling rate is Δt , then the second order temporal derivative of the wavefield can be estimated for each data point $u(x_i, y_j, t_k)$ by

$$\partial_t^2 u_{\omega'}(x_i, y_j, t_k) = \frac{u(x_i, y_j, t_{k-1}) - 2u(x_i, y_j, t_k) + u(x_i, y_j, t_{k+1})}{\Delta t^2} \tag{4.3}$$

and the second order spatial derivatives can be estimated by

$$\begin{aligned} \nabla^2 u_{\omega'}(x_i, y_j, t_k) = & \frac{u(x_{i-1}, y_j, t_k) - 2u(x_i, y_j, t_k) + u(x_{i+1}, y_j, t_k)}{\Delta x^2} + \\ & \frac{u(x_i, y_{j-1}, t_k) - 2u(x_i, y_j, t_k) + u(x_i, y_{j+1}, t_k)}{\Delta y^2} \end{aligned} \tag{4.4}$$

Here the second order spatial derivative is measured using 4 adjacent stations, two in the x direction and two in the y direction, to give a cross-shaped stencil. This means that it is not possible to calculate the derivatives for the stations at the boundary of the array and therefore results will not be produced close to the edge of the field area. Using the discrete approximation of the second order derivatives in Equations 4.3 and 4.4 it is possible to directly invert Equation 4.2 for the slowness s_M:

$$s_M(x_i, y_j) = \sqrt{\frac{\sum_{k=1}^{N_t} \mathrm{d}t \; \partial_t^2 u_{\omega'}(x_i, y_j, t_k) \; \nabla^2 u_{\omega'}(x_i, y_j, t_k)}{\sum_{k=1}^{N_t} \mathrm{d}t \; \partial_t^2 u_{\omega'}(x_i, y_j, t_k) \; \partial_t^2 u_{\omega'}(x_i, y_j, t_k)}} \tag{4.5}$$

where N_t is the number of time samples of the calculated second order gradients for each station.

The accuracy of the finite difference approximation decreases with a larger inter-station distance or shorter wavelength of the underlying function. Cao et al.

(In Press) analysed the approximation error and effect of noise on the results to give a relation between the calculated slowness s_M and the true slowness s_T

$$s_T = \gamma(s_T)\sqrt{1-\epsilon}s_M \tag{4.6}$$

where $\epsilon < 1$ is an unknown that depends on the signal-to-noise ratio of the spatial gradients and

$$\gamma(s_T) = \frac{\sqrt{|1-\cos 2\pi\Delta t|}}{\sqrt{|1-\cos s_T 2\pi\Delta x|}}\frac{\Delta x}{\Delta t}s_T \tag{4.7}$$

Using an estimate for ϵ, a fixed point iteration scheme is used to estimate s_T with $s_0 = s_M$ as a starting point so that

$$s_j = \gamma(s_{j-1})\sqrt{1-\epsilon}s_M \qquad j = 1, 2, ..., n \tag{4.8}$$

and the final $s_T \approx s_n$.

4.2.2 Neural Network Method

I forward modeled 100,000 randomly generated shear-wave depth-velocity structures for dispersion curves with frequencies 18, 20, 22 and 24 Hz using the DISPER80 subroutines by Saito (1988). The depth-velocity structures each have 11 layers that increase in thickness with depth (summarized in Table 4.1) since we expect our phase velocity data to provide lower resolution for deeper layers (Shapiro and Ritzwoller, 2002). For each layer in the velocity-depth structure the shear-wave velocities were uniformly randomly selected from a velocity range; for the top layer this was 50-300m/s since we know that there is a thick soil layer a priori, and for subsequent layers I used 300-2000m/s. An additional constraint was added to avoid unreasonably large velocity jumps between layers so that the velocity of the current layer v_i was in the range $v_{i-1} - 300m/s < v_i < v_{i-1} + 400m/s$ where v_{i-1} is the velocity of the layer directly above; if the velocity v_i lies outside of this range it was re-selected randomly. Once the dispersion curves had been generated for each structure, 5% Gaussian noise was added to the dispersion curves as it has been shown that this helps the network to generalise to new data (Meier et al., 2007b). The original depth structures were averaged vertically to

Velocity range in layer (m/s)	50-300	300-2000							
Original Layer thickness (m)	1	2	3	4	5	5	10	10	10	10	10
Up-scaled layer thickness (m)	3		7		10		10	20		-	

Table 4.1 Table summarising the parameterisation of shear velocity-depth structures used for training the MDN.

provide mean velocities in 5 depth layers with thickness 3m, 7m, 10m, 10m and 20m respectively (Table 4.1). I invert the data for velocities in these up-scaled depth structures so as to increase the degree of constraint on each layer offered by the measured dispersion. The up-scaled depth structures **m** and their associated dispersion curves **d** were used as the data-velocity structure pairs to train the MDNs. Separate MDNs were trained to output the distribution of shear velocity in each layer.

Multiple MDNs were trained for each layer and the network with the lowest cost value was selected as the final network for that layer. The outputs of each network are not true Bayesian posterior distributions because we do not know the absolute uncertainties on the phase velocity data but they do represent some measure of relative uncertainty on the shear velocity at each depth assuming that the data are all equally uncertain. The MDN's have the huge advantage over other nonlinear inversion methods such as Monte Carlo (Shapiro and Ritzwoller, 2002; Galetti et al., 2015; Zhang et al., 2018) of computational speed: once the networks have been trained they can be applied to new data sets in seconds (Meier et al., 2007b).

4.3 Field Data

Figure 4.1 shows the location of the field area which is farmland in south-east Edinburgh. Traffic from two main roads provides the main source of ambient seismic noise. Geophones were placed on a 8 by 11 grid with 5m spacing (Figure 4.1), the geophones had a 10 Hz corner frequency and recorded very little energy below 1 Hz. Ambient noise was recorded for just over 1 hour at a sampling rate of 125Hz. To attenuate amplitude spikes and account for varying amplitudes

between geophones an Automatic Gain Control (AGC) with a window length of 0.25s was applied and the complex spectrum was re-weighted by a applying a running avergae with a window size of 0.12Hz.

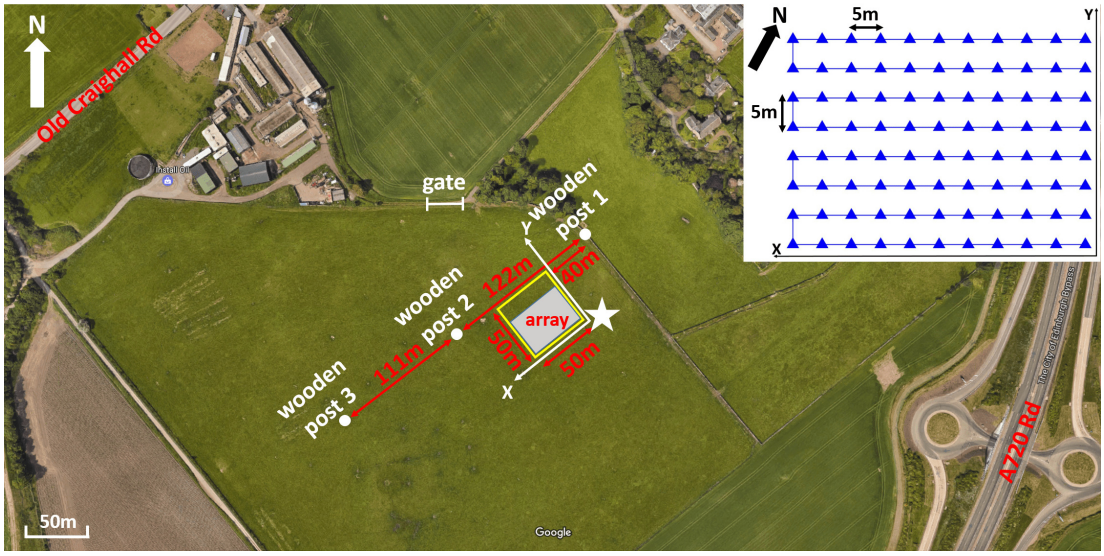

Figure 4.1 Satellite map of the acquisition field site south-east of Edinburgh. The white star represents the origin of the array. The two main roads (labelled red) which provide ambient noise are located at the northwest and southeast of the field site. The yellow square is the area of the array, geometry shown in the top right corner. From Cao et al. (In Press).

4.4 Results

4.4.1 Wavefield Gradiometry Results

The data was band-passed with a 5 Hz width Hann window centred every 1 Hz for frequencies in the range 3-36 Hz. The data was split into 11 separate 3 minute recordings and each recording was inverted using Equation 4.5 at each frequency to create multiple dispersion curves. Each dispersion curve is averaged over the whole array to create 11 average dispersion curves. Since the finite difference approximations of Equation 4.5 cause errors in the final calculated slowness, the average dispersion curves are compared to the frequency-wavenumber spectrum to determine the optimum value for ϵ. Figure 4.2 shows the frequency-wavenumber

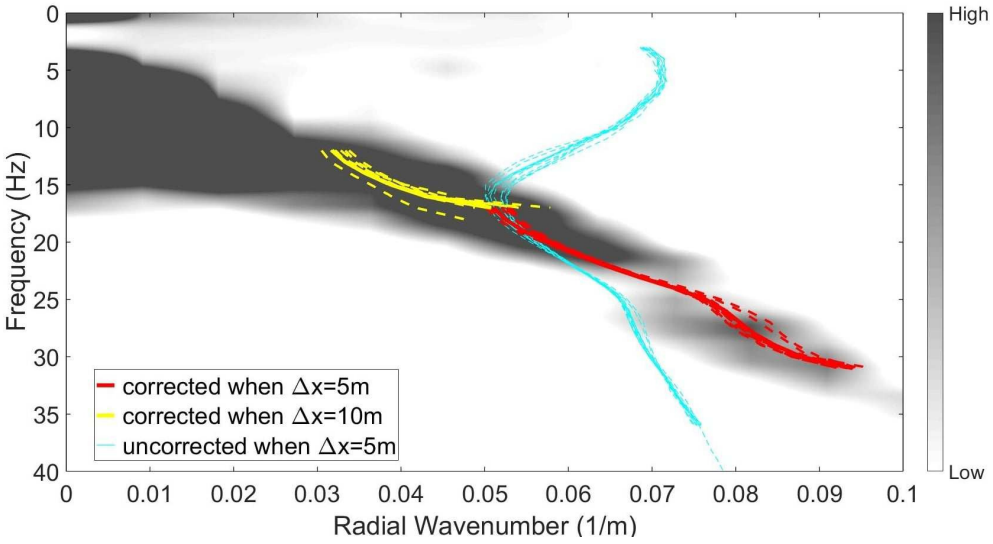

Figure 4.2 Frequency-wavenumber spectrum (background grey shading) and phase slowness curves from gradiometry before (blue) and after (red and yellow) correction using Equation 4.8. Dotted lines show the 11 average dispersion curves and the solid line is the combination of the dotted lines in each case. The correction is only applied to frequencies above 12Hz. From Cao et al. (In Press)

spectrum with the original dispersion curves in blue. Equation 4.7 is simplified to account for only spatial finite difference approximation errors to give

$$\gamma_2(s_T) = \frac{s_T 2\pi |f| \Delta x}{\sqrt{2|1 - \cos s_T 2\pi f \Delta x|}} \tag{4.9}$$

This equation is valid for equally spaced arrays i.e. $\Delta x = \Delta y$. Using Equations 4.6 and 4.9 an optimal value of $\epsilon = 0.2$ was determined to give the best agreement between the spatially averaged dispersion curves and the frequency-wavenumber spectrum above 15 Hz (Red lines in Figure 4.2). Below 15 Hz the correction does not work for the given value of $\epsilon = 0.2$, however if the inter-station spacing is increased to 10m the dispersion curves can be corrected to agree with the frequency-wavenumber spectrum as shown by the yellow line in Figure 4.2.

To produce the final field data results the central 30 minutes of the hour-long recording are processed, and using Equations 4.5, 4.6, 4.9 with $\epsilon = 0.2$, phase velocity maps are produced at 18 Hz, 20 Hz, 22 Hz, 24 Hz (4.3).

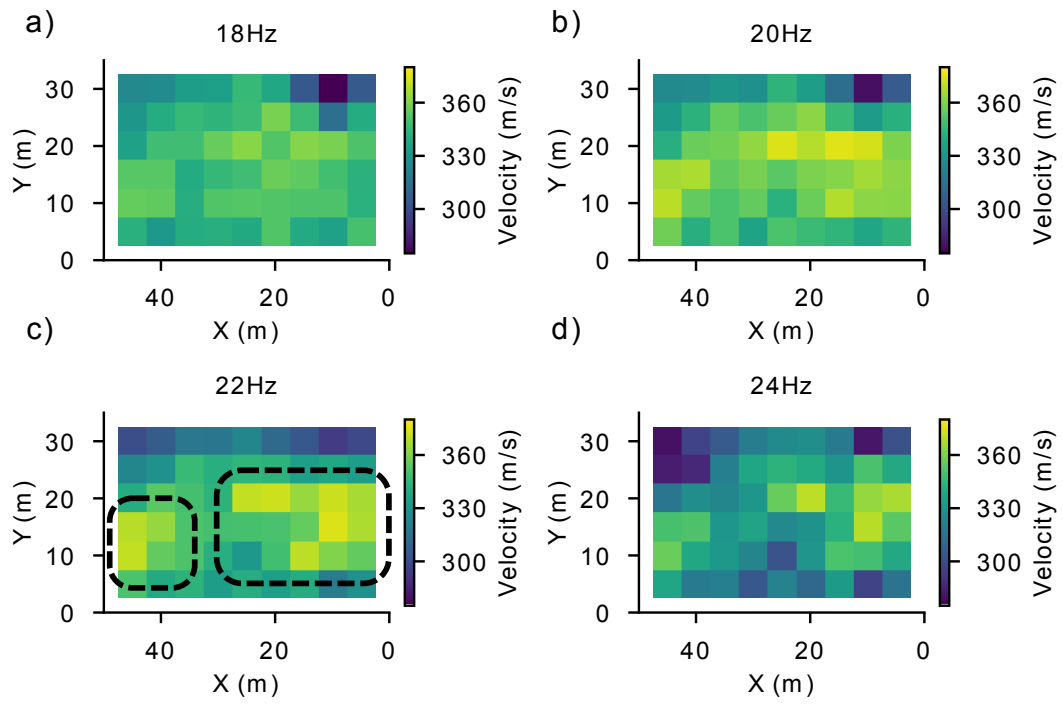

Figure 4.3 Phase velocity maps obtained from the ambient-noise field data set by gradiometry at four centre frequencies: (a) 18 Hz, (b) 20 Hz, (c) 22Hz, (d) 24Hz. The axis limits correspond to the complete survey dimensions; the lack of velocity information near the edges is due to the finite difference method requiring four neighbouring stations.

4.4.2 Neural Network Results

Using the phase-velocity maps for frequencies 18, 20, 22 and 24 Hz shown in Figure 4.3 the 4-point dispersion curve was extracted at each grid location and inverted for a 1D depth-velocity structure beneath that location. Figures 4.4 and 4.5 are maps of the mean result and the corresponding standard deviation of the output probability distribution at each depth level. The top two depth layers in the mean maps (Figure 4.4a and 4.4b) show a good correlation with the phase velocity maps in Figure 4.3 at 22Hz shown by the black dotted box.

To assess to accuracy of the MDN result I ran a 1D Markov chain Monte Carlo (McMC) depth inversion beneath each point in the grid. I used a fully non-linear Markov chain inversion method as described in Galetti et al. (2017) based on the code from Bodin and Sambridge (2009) which returns a set of solutions sampled

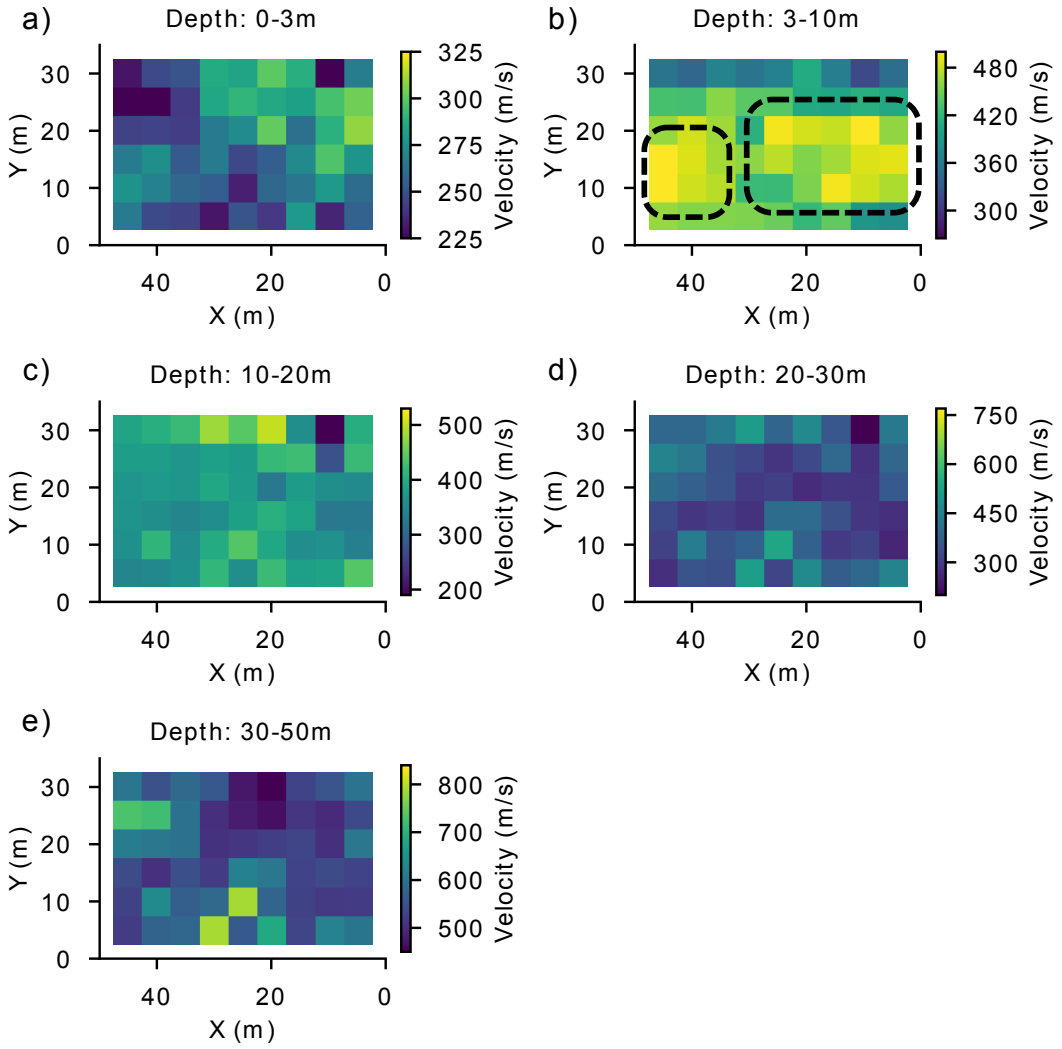

Figure 4.4 The mean shear velocity maps from MDN inversion for a) 0-3 m, b) 3-10 m, c) 10-20 m, d) 20-30 m, e) 30-50 m.

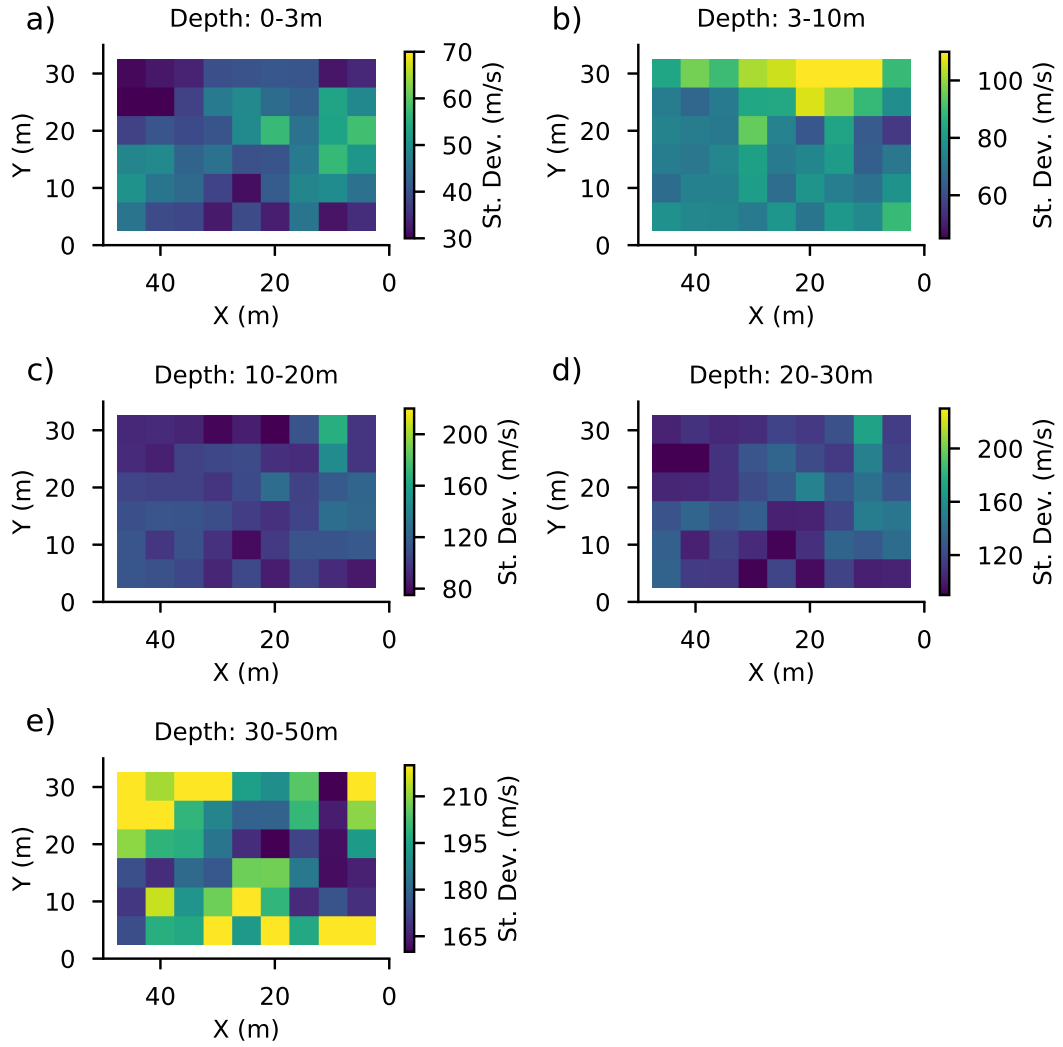

Figure 4.5 The standard deviation of the shear velocity maps from MDN inversion for a) 0-3 m, b) 3-10 m, c) 10-20 m, d) 20-30 m, e) 30-50 m.

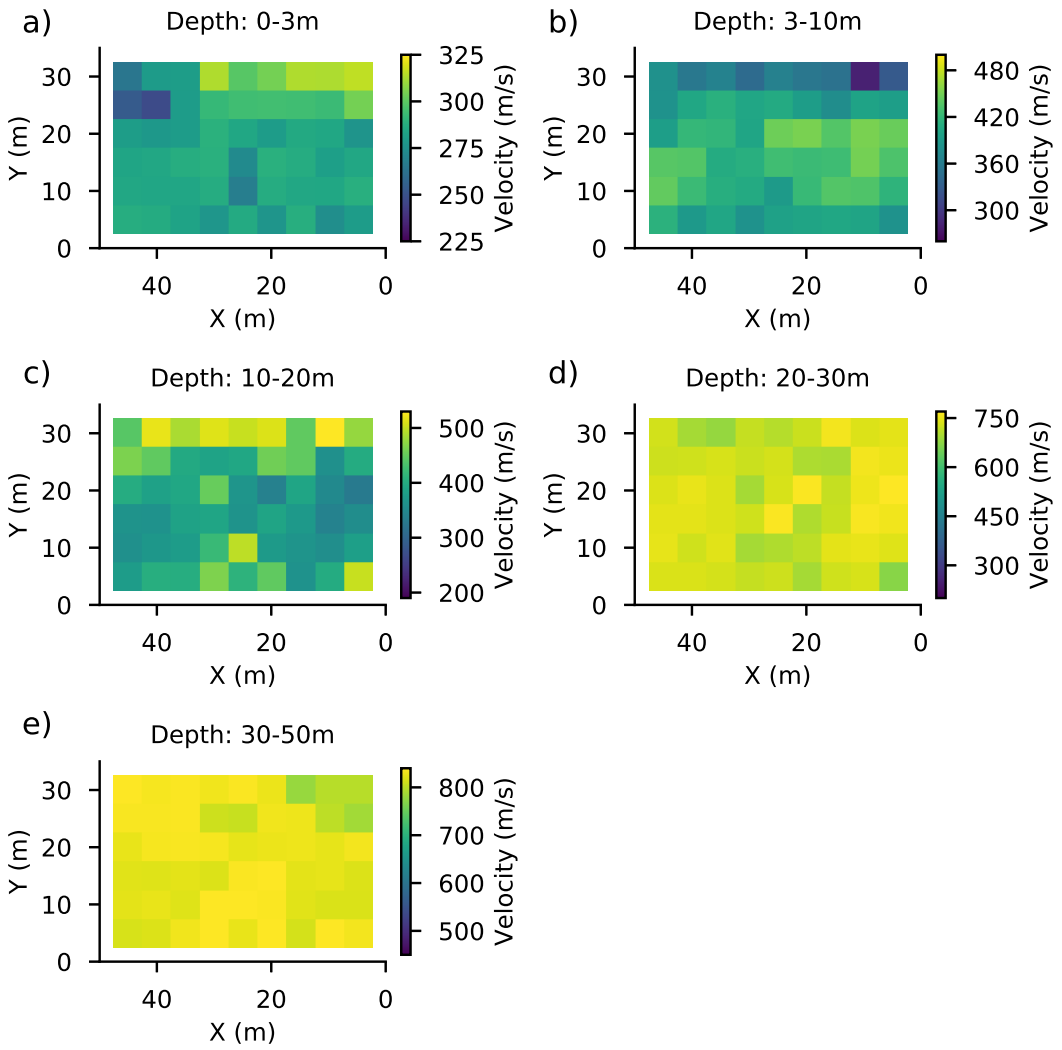

Figure 4.6 The mean shear velocity maps from Monte Carlo inversion for a) 0-3 m, b) 3-10 m, c) 10-20 m, d) 20-30 m, e) 30-50 m.

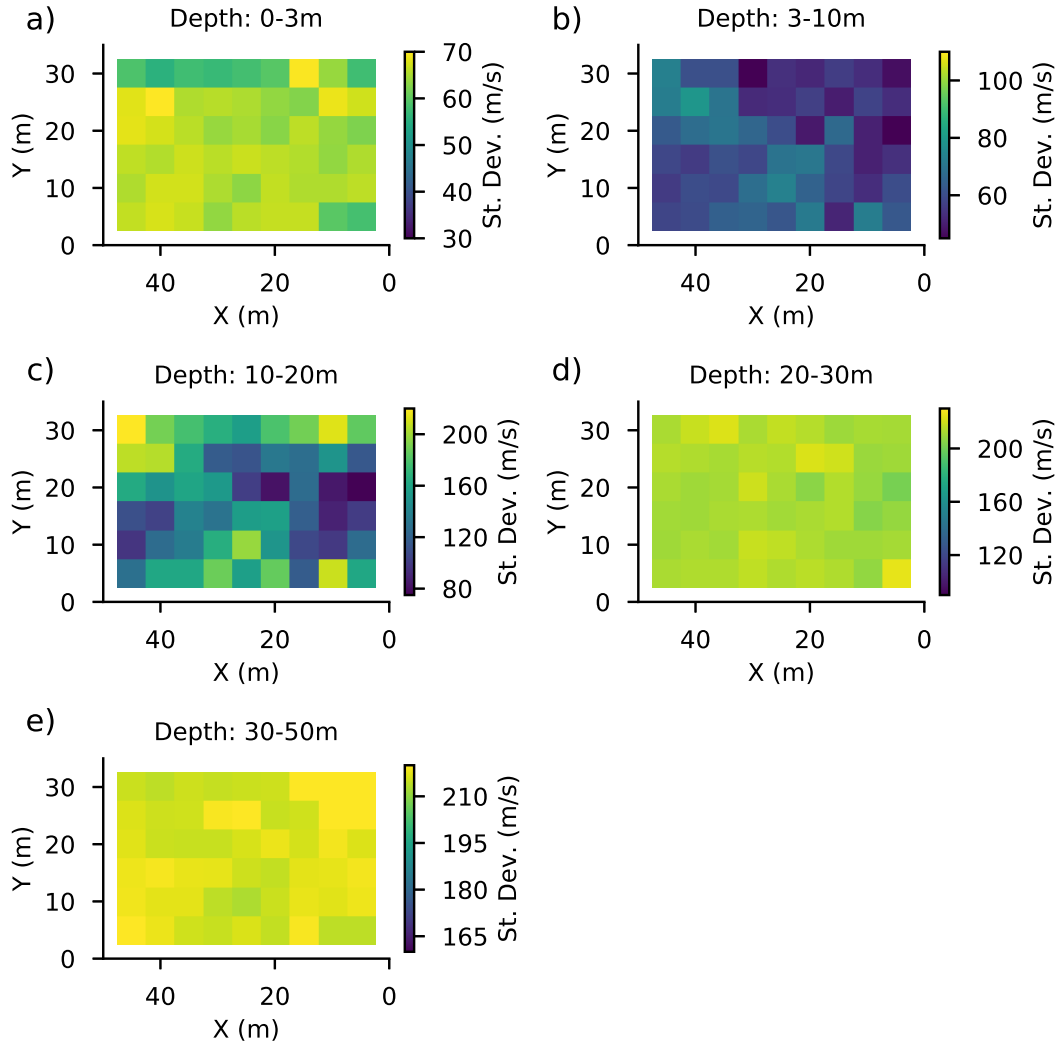

Figure 4.7 The standard deviation of the shear velocity maps from Monte Carlo inversion for a) 0-3 m, b) 3-10 m, c) 10-20 m, d) 20-30 m, e) 30-50 m.

from the Bayesian posterior probability density function. To obtain a comparable solution to the network based results, the number and the depth of layers in the McMC were fixed to be the same as those used to train the MDN. The maps in Figures 4.6 and 4.7 show the mean and standard deviation respectively of the McMC results. The depth layers down to 20m in Figure 4.6 show a structural similarity with the MDN inversion method in Figure 4.4, however the velocities are generally higher in the Monte Carlo inversion. For the deeper layers (>20m) the mean and the standard deviation of the McMC result (Figures 4.6 and 4.7, d and e) are much higher than the MDN inversion and also show very little structure in common. It appears that all points at this depth converged on a similar solution and no information about the structures are recovered at this depth, indicating that the frequencies inverted are only sensitive to a depth of 20m.

As an example, the 1D depth profile for location (x,y) = (40 m, 10 m) is shown in Figure 4.8a and the individual probability density functions from MDN and McMC inversion at each depth level are shown in Figure 4.8b-f. The MDN inversion results are consistent with the McMC results down to a depth of 20m. Below this the Monte Carlo and MDN solution are quite different.

4.5 Discussion

Using the scalar wave equation, frequency dependent surface wave velocity maps were produced from ambient noise. From these, depth maps were produced for shear-wave velocity and an indication of their associated uncertainties were produced from inversions by MDNs. The mean velocity maps (Figure 4.4) show a correlation with the phase velocity maps produced in Figure 4.3.

4.5.1 Monte Carlo comparison

A Monte Carlo sampling solution was performed as a comparison to the MDN inversion method. At a depth of less than 20m the MDN inversion and McMC solution show similar spatial velocity structure. The 3-10m depth map (Figures 4.4b and 4.6b) show a simliar structure but the absolute velocities are different,

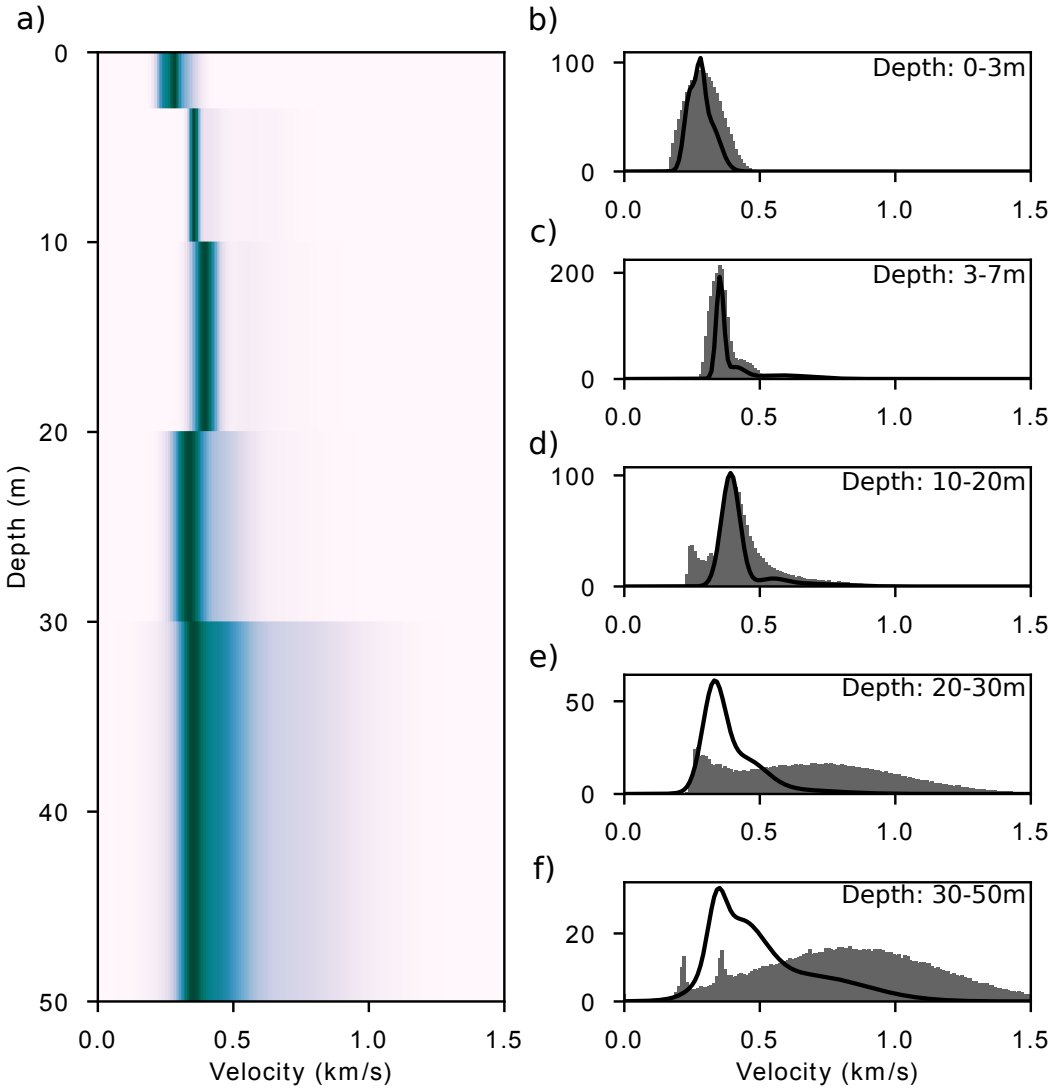

Figure 4.8 1D depth inversion result for location (x,y) = (40 m, 10 m). a) The MDN normalised posterior density function result: dark colors represent areas of higher probability. The posterior density function at each depth level is shown in panels b) 0-3 m, c) 3-10 m, d) 10-20 m, e) 20-30 m, f) 30-50 m. The black line represents the MDN pdf, the grey shaded area is the histogram from the Monte Carlo inversion.

however the difference in velocity falls with the range of uncertainty shown in Figures 4.5b and 4.7. Below 20m the velocity and uncertainty of the McMC solution is much higher than MDN inversion. The lack of structure and high uncertainty values in the McMC solution suggest that the data does not give any information for the velocity structure, and the MDN appears to give poor estimates of the posterior at this depth.

4.5.2 Inversion Speed

Standard nonlinear 3D inversion inversion methods, such as Monte Carlo methods (Bodin and Sambridge, 2009; Galetti et al., 2015; Zhang et al., 2018), that are commonly used to produce probabilistic results are computationally expensive. By contrast, a neural network takes a couple of hours to train and can then be used repeatedly, without further training, to produce results in a matter of seconds for any data provided as input. In this study it took less than one second to invert the 54 dispersion curves from the gradiometry result to create Figures 4.4 and 4.5. However, this does not include the total time needed for training a network. In this example 5 networks were trained, one for each depth layer and each network took on average 18 CPU minutes to train, 1.5 CPU hours in total. The Monte Carlo results in Figures 4.6 and 4.7 required 22 CPU minutes per point, approximately 20 CPU hours for the full area. Even for this small example, the MDN is therefore more computationally efficient (including training time) than the Monte Carlo sampling. If this method were used over a larger area with a higher number of data locations, the MDN method could be several orders of magnitude faster than McMC. For repeated inversions such as monitoring of a field area, the MDN does not need to be retrained so future inversions of the same field area are essentially free.

4.5.3 Near real time monitoring

The gradiometry and wave equation inversion methods in Section 4.2 are computationally cheap. For a small area such as the results in this chapter the sequence takes no longer than 10-20 seconds, from the pre-processing of the data to the

phase velocity maps. Since the processing of the data is so rapid the determining factor for how long the method takes is the time needed to record the ambient noise. The final results used 30 minute length ambient noise records, however very short noise records (3 minutes) were used to determine the value of ϵ shown in Figure 4.2 and these short recordings could potentially be used to also produce phase velocity maps. Since the neural network inversion takes only one second once the network is trained, the whole sequence from acquisition to depth inversion could take less than 4 minutes.

This rapid method could be used in two different ways. Either a spatially small acquisition can be used as a 'rolling' array, where the equipment is moved around a larger area of interest (the back line of geophones being iteratively moved to the front). With a short recording time the array can be installed and moved as fast as possible, and still we would record for the time needed to invert the data. Thus the surface wave velocity structure could be estimated efficiently for the entire area. Alternatively, for a larger networks of sensors such as permanent reservoir monitoring systems over 10's of kilometers, where data is continuously recorded, the inversion method can be used to monitor the subsurface structures in the field. The inversion process will now take 2-3 minutes due to the increase in the volume of data however for monitoring purposes this is a relatively short length of time that may for example produce hourly 3D tomographic subsurface snapshots.

4.6 Conclusion

This chapter presented a method for rapid probabilistic inversion for 3D seismic depth-velocity structures using wavefield gradiometry, wave equation inversion and neural networks. Only short (30 mins) ambient noise recordings are needed to invert the dispersive scalar wave equation using measured second order derivatives of the wavefield. These phase velocity maps can then be inverted for 3D seismic depth-velocity structures using mixture density networks, with uncertainty estimates. The full inversion process from field data to 3D depth-velocity structure is computationally cheap, opening up the possibility of near-real time monitoring using dense arrays.

Chapter 5

Probabilistic Neural-Network Based 2D Traveltime Tomography

Travel time tomography for the velocity or slowness structure of a medium is a highly non-linear and non-unique inverse problem. Monte Carlo methods are becoming increasingly common choices to provide probabilistic solutions to tomographic problems but those methods are computationally expensive. Neural networks can often be used to solve highly non-linear problems at a much lower computational cost when multiple inversions are needed from similar types of data. I present the first method to perform fully non-linear, rapid and probabilistic Bayesian inversion of travel time data for 2D velocity maps using a form of neural network called a mixture density network. I compare multiple methods to estimate probability density functions that represent the tomographic solution, using different sets of prior information and different training methodologies. I demonstrate the importance of prior information in such high dimensional inverse problems due to the curse of dimensionality: unrealistically informative prior probability distributions may result in better estimates of the mean velocity structure, however the uncertainties represented in the posterior probability density functions then contain less information than is obtained when using a less informative prior. This is illustrated by the emergence of uncertainty loops in posterior standard deviation maps when inverting travel time data using a less informative prior, which are not observed when using networks trained on

prior information that includes (unrealistic) a priori smoothness constraints in the velocity models. I show that after an expensive program of training the networks, repeated high-dimensional, probabilistic tomography is possible on timescales of the order of a second on a standard desktop computer.

5.1 Introduction

Seismic travel time tomography is often used to reconstruct images of the interior of the Earth (Aki et al., 1977; Dziewonski and Woodhouse, 1987; Montelli et al., 2004; Shapiro et al., 2005), but is a significantly non-linear and non-unique inverse problem. To find solutions with minimal computation, the physics relating local wave speed to measured travel times is usually simplified by linearization (Rawlinson et al., 2010), but this creates large differences between linearized and true probabilistic solutions (Galetti et al., 2015). Increases in compute power now allow fully nonlinear Monte Carlo sampling solutions to be found without linearization, to solve problems in 2D (Bodin and Sambridge, 2009; Galetti et al., 2015) and 3D (Hawkins and Sambridge, 2015; Piana Agostinetti et al., 2015; Zhang et al., 2018, 2019). Using Bayesian methods, such solutions provide samples (example tomographic models) that fit the data to within their measurement uncertainties, are consistent with available prior information, and are distributed according to the posterior probability density function (pdf) across the parameter space; this pdf constitutes the full solution of tomographic problems. Nevertheless, such solutions are acquired at significant expense, typically requiring weeks of compute time for realistic data sets and expensive storage of large sample sets.

An alternative approach to estimate the posterior pdf is to use prior sampling (Devilee et al., 1999; Käufl et al., 2016). In this case samples are created before inference using only available prior knowledge. The set of samples can then be interrogated for examples that are consistent with any particular data set (a method called *resampling* (Sambridge, 1999)) or used to parametrise a function that relates data to models which can then be used to solve the inverse or inference problem (Roth and Tarantola, 1994).

In this work I use a neural network-based method to perform the inversion. Neural networks (NNs) can approximate any nonlinear relationship between two parameter spaces, given a so-called training set of example pairs of dependent and independent parameter values under that relationship (Bishop, 1995). In travel time tomography the forward solution is known and calculable, but the inverse solution is highly non-linear and non-unique. In such cases the forward computation can be used to create the prior set of samples known as a *training set*, of random models drawn from the prior pdf; these can be used to train the neural networks to approximate the inverse mapping. The prior samples are only needed during the training process which needs only to be performed once - thereafter NNs can be evaluated relatively efficiently. This allows the inference step to be run rapidly for any new data set on standard desktop computers, and the overall cost of the method per tomographic problem decreases rapidly with the number of problems to be solved.

Neural network-based inversion methods have been applied to various tomography problems in the past. Roth and Tarantola (1994) first used NNs to estimate subsurface velocity structure from active source seismic waveforms, Moya and Irikura (2010) performed velocity inversion with a neural network using waveform data from earthquakes, and Araya-Polo et al. (2018) used semblance gathers as input to a network to invert for velocity structure. Gupta et al. (2018) used a convolutional network to learn an ensemble of simpler mappings in a low-dimensional space before reconstructing the image by combining the mappings. Dictionary learning methods (Mairal et al., 2014) create sparse representations of the data and can be used to create a set of representations of features. Bianco and Gerstoft (2018) performed linearized 2D surface wave travel time tomography using dictionary learning to regularise the inversion.

The methods mentioned above and in Kong et al. (2018) all provide only deterministic solutions to the inversion. Since the solution to tomographic problems is always non-unique, in order to assess the worth of any model estimate I require that neural networks produce full probabilistic information about the set of models in the inverse problem solution (the posterior pdf). Devilee et al. (1999) solved the first probabilistic geophysical inverse problem using NNs. They proposed a

variety of methods to train NNs to provide discretised Bayesian posterior pdfs. Mixture density networks (MDNs) are a class of augmented neural networks that output a probability distribution that is defined as a sum of analytic pdf kernels such as Gaussians (Bishop, 1995). MDNs can be trained such that for any input data this distribution approximates the posterior pdf. These methods have been used at a global scale to invert surface wave velocities for global crustal thicknesses and seismic velocities (Meier et al., 2007a,b) and for water content in the mantle transition zone (Meier et al., 2009), at a reservoir scale to infer petrophysical parameters from velocities (Shahraeeni and Curtis, 2011; Shahraeeni et al., 2012), for earthquake source parameter estimation (Käufl et al., 2014, 2015) and to assess the uncertainty in model parameters of the Earth's global average (1-dimensional) radial velocity structure from P-wave travel time curves (De Wit et al., 2013). They have also been used in conjunction with Markov random fields and other statistical and graphical models to solve geophysical inverse problems with spatially sophisticated prior information (Nawaz and Curtis, 2017, 2018, 2019). They have been used in conjunction with seismic gradiometry to perform near-real time 3D surface wave tomography (Cao et al., In Press). These studies demonstrated that the pdf obtained from an MDN is comparable to a Monte Carlo sampling solution but is obtained at much lower computational cost in the cases where similar inverse problems must be solved repeatedly with different data sets, and that at the moment of application MDNs provide probabilistic solutions almost instantaneously.

I show for the first time that MDNs can perform fully non-linear, rapid and probabilistic 2D tomography from travel time data. I compare different methods for creating the prior training set and performing the neural network inversion. The networks create approximate mean velocity models and estimates of the full marginal posterior pdf's, virtually instantaneously. Thus, in return for accepting approximate posterior pdfs I obtain a significant computational saving compared to Monte Carlo methods.

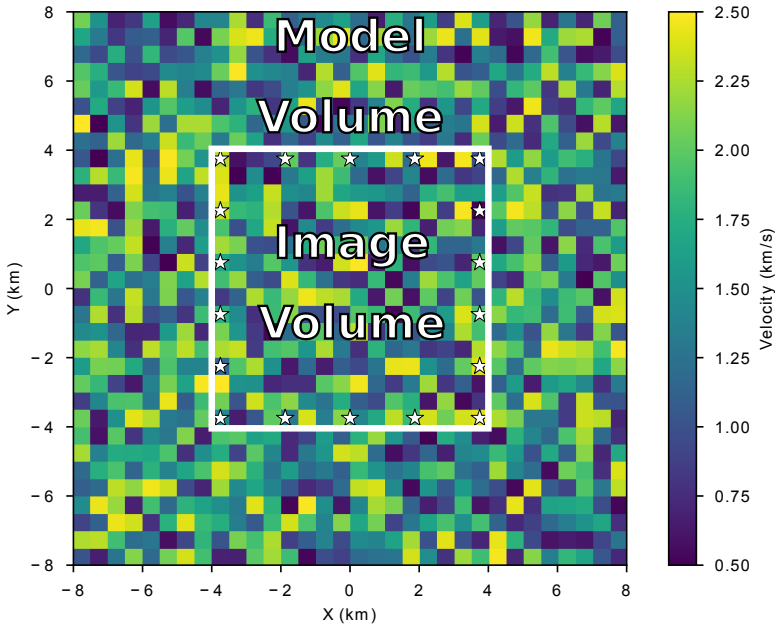

Figure 5.1 Geometry of velocity models. Larger model with limits (-8,8) in the X and Y direction is the *Model Volume* within which the travel-times are calculated. The smaller model bounded by a white box with limits (-4,4) in the X and Y direction is the *Image Volume* which I wish to image. White stars represent the location of co-located sources and receivers, between which travel time data are obtained.

5.2 Method

5.2.1 Model Parametrisation and Traveltime Data

I define the geometry of the tomography problem to be that shown in Figure 5.1. I fix the locations of 18 wave energy sources and receivers (shown in Figure 5.2), and parametrise the wave speed or velocity across the *Model Volume* within which the forward relationship predicts travel times of the first arriving energy between any source-receiver pair. Travel times $\mathbf{d}_i$ between all possible source-receiver pairs are calculated using an eikonal raytracer (Rawlinson and Sambridge, 2004, 2005). The traveltimes from the 4 velocity models shown in Figure 5.2 are shown in Figure 5.3. Such travel times are used herein to image the velocity structure within the smaller *Image Volume* - wave speeds outside of that area are disregarded and thus constitute nuisance parameters. I use a larger volume to calculate the forward relationship to avoid raypaths travelling along the boundary of the model and causing misleading travel times.

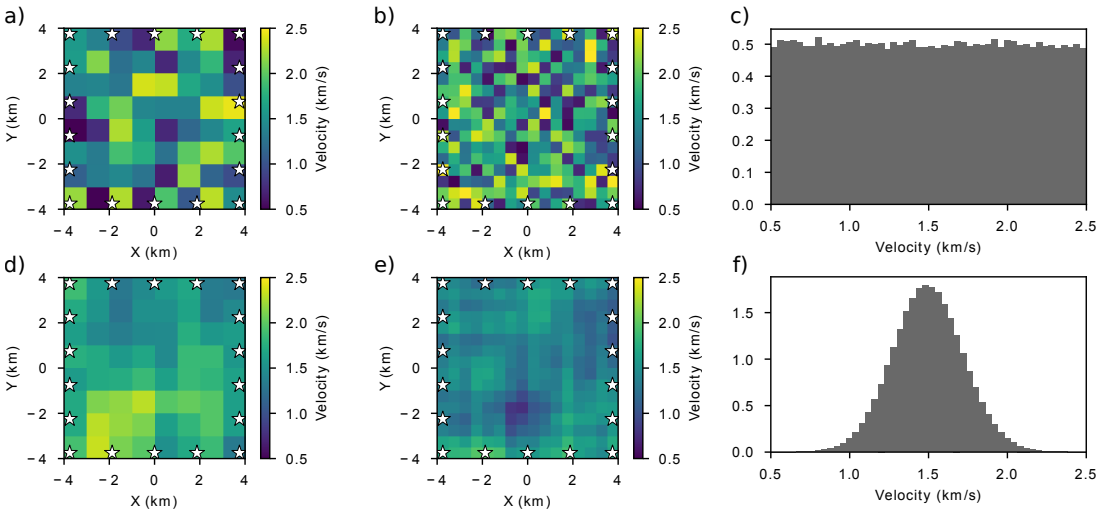

Figure 5.2 Example velocity models from the 4 training sets that are randomly selected from Uniform distributions on an (a) 8 by 8 grid and (b) 16 by 16 grid, or are randomly selected and then smoothed with a spatial averaging filter on a (d) 8 by 8 grid and (e) 16 by 16 grid. White stars represent the location of co-located sources and receivers. The prior distribution of the training set is shown for one cell in the model given a fixed neighbouring cell for (c) models selected from a Uniform random distribution and (f) similar models after spatial smoothing.

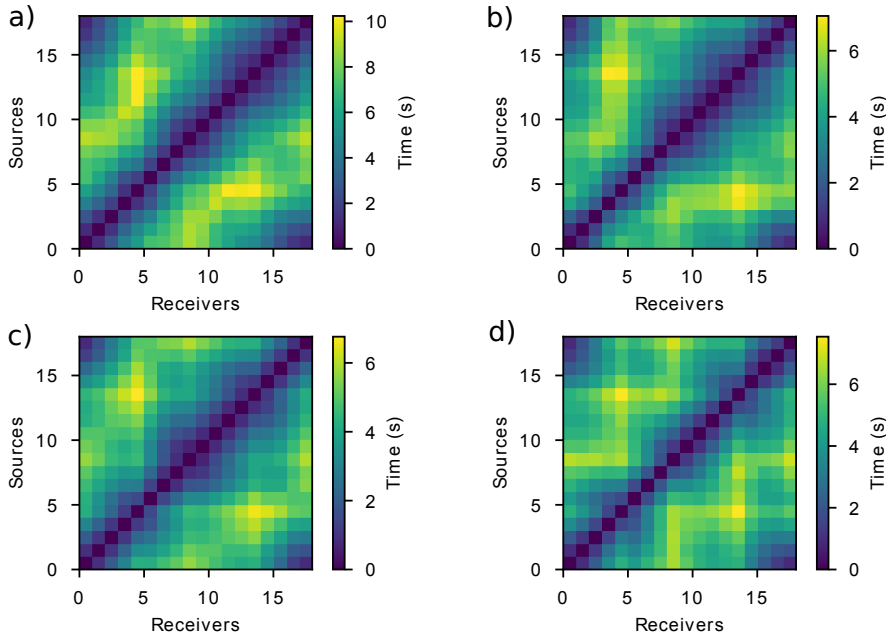

Figure 5.3 Corresponding data from the four velocity models in Figure 5.2 that are randomly selected from Uniform distributions on an (a) 8 by 8 grid and (b) 16 by 16 grid, or are randomly selected and then smoothed with an averaging filter on a (c) 8 by 8 grid and (d) 16 by 16 grid.

I construct four separate training sets, each of 2.5 million discretised models where each model represents a 2D heterogeneous velocity structure. Two of these training sets are created on an 8 x 8 coarser grid of cells and two are created on a 16 x 16 finer grid of cells within the *Image Volume* (and the same resolution extends throughout the *Model Volume*). Each of the four datasets is created by selecting a random wave speed in each cell independently from the Uniform prior distribution $U(0.5km/s, 2.5km/s)$. All models in one finer data set and one coarser data set are then smoothed using a 2D averaging filter window which was square of size 5x5 cells for the finer model and 3x3 cells for the coarser model. Thereafter the velocities are normalised to the same absolute range as the original random models for ease of comparison of results. Then, the travel times between all source-receiver pairs are calculated for all models, in all four training sets (examples are given in Figure 5.3).

With this method I create training sets with two different amounts and types of prior information. The two sets of random unsmoothed velocity models have relatively weak prior information with no correlations between neighbouring cells. This has the advantage that any type of velocity contrast between neighbouring cells would be consistent with the prior pdf and hence can in principle be imaged using the associate trained network given sufficiently informative data (see below). This is demonstrated by the uniform distribution of the histogram in Figure 5.2c which shows the probability of the velocity of the adjacent cell given that the velocity of the central cell is 1.5km/s. On the other hand this implies that the prior pdf is Uniform over a 64- and 256- dimensional space for the coarser and finer training sets respectively; these spaces are therefore extremely sparsely sampled by the 2,500,000 training set models due to the curse of dimensionality (Curtis and Lomax, 2001). This implies that over most of these two spaces the prior pdf is entirely unrepresented by 'proximal' samples.

The two sets of smoothed velocity models embody stronger prior information as the speeds in neighbouring cells are correlated. This is demonstrated in Figure 5.2f where the distribution of possible velocities in adjacent cells given that the velocity of the central cell is 1.5km/s is approximately Gaussian. This means that models with larger velocity contrasts between neighbouring cells are not

represented in the training data set and hence will be precluded from inversion results. This may or may not be advantageous depending on the true prior information about the form of the structure being imaged. However, it has the advantage that the effective space (manifold) of models consistent with the prior information is considerably smaller than that for the unsmoothed models, so that the finite-sized training set may better represent the form of the prior pdf.

5.2.2 Network Configurations

I train separate MDNs to predict the marginal probability distribution $p(m^i \mid \mathbf{d})$ of velocity m^i in cell i in each of the two sizes of models. For each cell in the model an ensemble of MDNs predicts the velocity at that location from a network that has been trained the full traveltime information from all sources and receivers and the true velocity of that cell alone. For the finer datasets I train 4 MDNs and for the coarser datasets I train 8 MDNs at each location i. I use different configurations as well as randomly initialised internal network parameters (commonly referred to as weights and biases) for each network because diversity in the ensemble generally leads to better predictions (Dietterich, 2000). Appendix A outlines the different network configurations. For each network I use a Gaussian mixture consisting of 15 kernels. The precise number of kernels is not important as long as it is larger than the number required to represent the marginal posterior pdf in each model cell. The network can either reduce the amplitude of the mixture parameter α_{ij} to close to zero to remove unnecessary kernels, or can combine unnecessary kernels by giving them a similar μ and σ to other kernels (Bishop, 1995). In practice I found the maximum number of kernels with significant weight used in any mixture was 8.

I also train networks to invert for the full model (velocities in all cells at once) using a single network. In this case I use a convolutional network with 3 convolutional layers followed by 3 fully connected layers and 15 kernels for the Gaussian mixture. I train 10 networks with 5 different network configurations (each configuration is trained twice with random weight initialisation). Layer sizes were selected using the python library hyperopt (Bergstra et al., 2015) and

Appendix C gives further description of the networks used. The same network configurations were trained on all four training sets.

For every training run for each network configuration I use 85% of the training dataset to train the network, 10% of the dataset as a validation set during training, and 5% as a test set to evaluate the final network once training has finished. The training set is used in the optimisation of network parameters. The parameters are updated iteratively so that the output of the network best represents the training set sample distribution. To avoid over-fitting the network to the data the cost function is also periodically evaluated over the validation set; when the error on the validation set stops decreasing I end the training optimisation. Once all of the networks have been trained I evaluate the final network performance using the test set and sum the networks across the ensemble using equation 2.16.

5.3 Results

5.3.1 Result Evaluation

I tested the trained networks by applying them to synthetic data sets calculated for velocity models created specifically to test the performance of each type of network. The quality of the mean of the inverted probability distributions of 2D velocity models (comprising 1D marginal posterior pdfs in each model cell in the cases where networks were trained for each cell individually) are compared against the true velocity model using the structural similarity index metric (SSIM). This metric is based on 3 relatively independent comparison measurements: luminance, contrast and structure (Appendix D). SSIM can assume values between -1 and 1: a value of 1 indicates the images are identical, 0 indicates no structural similarity and negative values occur when local structure is inverted. SSIM differs from other quality indicators such as mean squared error (MSE) in that it measures the quality of an image in structure and pixel value compared to a ground truth, rather than the absolute squared errors (which often do not mean much to someone who is trying to interpret the resulting images).

I compare the information gain between the prior $p(\mathbf{m})$ and the posterior $p(\mathbf{m}|\mathbf{d})$ distribution using the Kullback-Leibler (KL) divergence which measures the difference between two probability distributions. It is the information content of the prior weighted by the posterior, known as cross entropy, minus the entropy of the posterior. The measure is given by the expectation of the log difference between the probability of data in the posterior with the prior so that

$$D_{\mathrm{KL}}\left(p(\mathbf{m}|\mathbf{d}), p(\mathbf{m})\right) = \int_{-\infty}^{\infty} p(\mathbf{m}|\mathbf{d}) ln\left(\frac{p(\mathbf{m}|\mathbf{d})}{p(\mathbf{m})}\right) dx \tag{5.1}$$

where a higher D_{KL} indicates that the posterior pdf has gained information over the prior and $D_{\mathrm{KL}} = 0$ occurs when the two distributions are the same. This can be used as an indication of the effectiveness of the network: if D_{KL} is close to 0 then the network has been able to learn little, if anything at all, from the data.

5.3.2 Prior

To show the effect of the prior on the models I inverted synthetic data for the three velocity models shown in Figures 5.4a and 5.5a using networks trained with weak prior information (unsmoothed training models) in Figure 5.4 and those trained with stronger prior information (smoothed models) in Figure 5.5. The test models were defined on a grid finer than the training sets on a 32x32 grid, which is finer than either of the training sets; this ensures that I evaluate the networks using models that are outside the range of those used for training. For all test models it is clear that with stronger prior information the networks better resolve the velocity structure, shown generally by the much higher SSIM values in Figure 5.5b and 5.5c compared with the corresponding values in Figure 5.4b and 5.4c. This is true even though the test models contradict the stronger prior information: they all contain structures that are not smooth.

The velocity model in the left-hand column has a background velocity (cells surrounding the central anomaly) equal to the mean of the prior pdf and a circular low velocity, and is estimated well in both inversions using weaker prior information training sets (Figure 5.4). However, even a small increase in complexity in velocity models gives poor inversion results as shown by the central column of velocity

models. For these, all the velocities are increased compared to the left column, and in particular the background velocity is increased away from the mean of the prior. In this case the networks with weaker prior information are unable to recover much, if any, of the true structure. If stronger prior information is included in the training set the networks accurately predict a larger variety of velocity models. The true structures of the two circular models in Figure 5.5 are closely reproduced in the inversion. Sharp contrasts in velocity in the true model are translated to more gradual changes in velocity in the estimates (for both grid sizes) due to the smoothness in the prior pdf. Despite this, the SSIM values show that results are very well correlated with the true model. For the more geologically reasonable model in the right column of Figure 5.5 which includes a structure that might be generated by a fault, networks trained using stronger prior information on both grid sizes produce models that are nearly identical to the true model. Even though the true model contains a sharp contrast boundary, the inverted models still contain a (slightly smoother) version and the overall structure of the true image is maintained.

The effect of stronger prior information is shown in the posterior pdfs in Figure 5.6. I display the posterior marginal pdfs at three locations indicated in the upper right hand model in Figure 5.4a: a location in the high velocity zone (triangle), the low velocity zone (circle), and at the edge of the sharp contrast where the inversion struggles to image correctly (star). The KL divergence values are shown above the corresponding posterior marginal pdf. The most striking feature is the much higher KL values for the networks trained with the stronger prior information (rows b and d) indicating a larger information gain in the posterior pdf compared to the prior pdf than is obtained when training with Uniformly random models. In fact, the low KL values for the latter cases imply that nearly no information was gained from the data, and even though a rough approximation of the mean can be found the uncertainties on those values remain large.

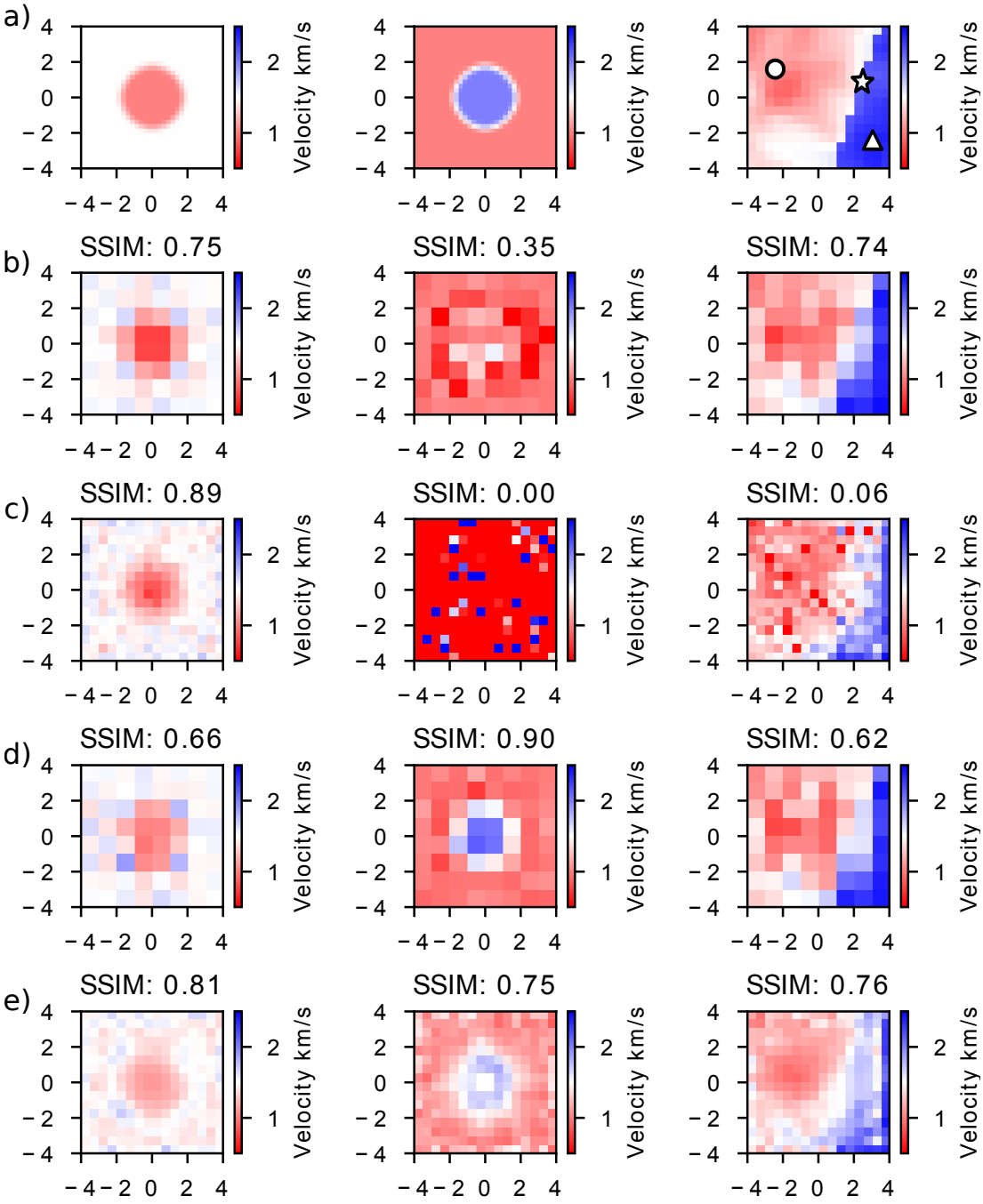

Figure 5.4 (a) True velocity models. Using a randomly generated training set drawn from a Uniform distribution, mean velocities from separate-cell MDN inversions for (b) an 8 x 8 model and (c) a 16 x 16 model, and from full-model MDN inversions for (d) an 8 x 8 model and (e) a 16 x 16 model. The corresponding SSIM values are shown above each result (see Appendix D for definition of SSIM).

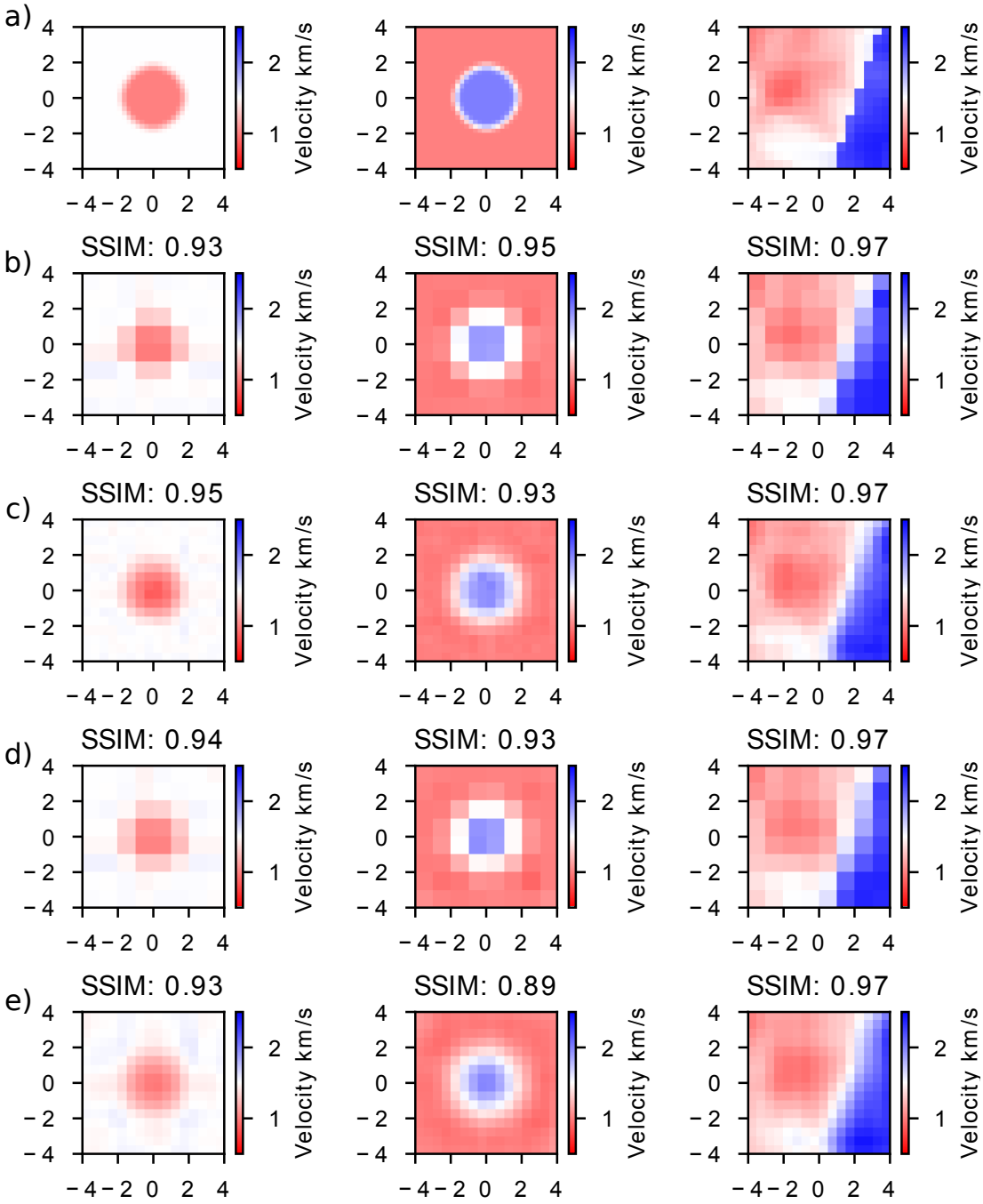

Figure 5.5 (a) True velocity models. Using a training set with spatially smoothed velocities, mean velocities from separate-cell MDN inversions for (b) an 8 x 8 model and (c) a 16 x 16 model, and from full-model MDN inversions for (d) an 8 x 8 model and (e) a 16 x 16 model. The corresponding SSIM values are shown above each result (see Appendix D for definition of SSIM).

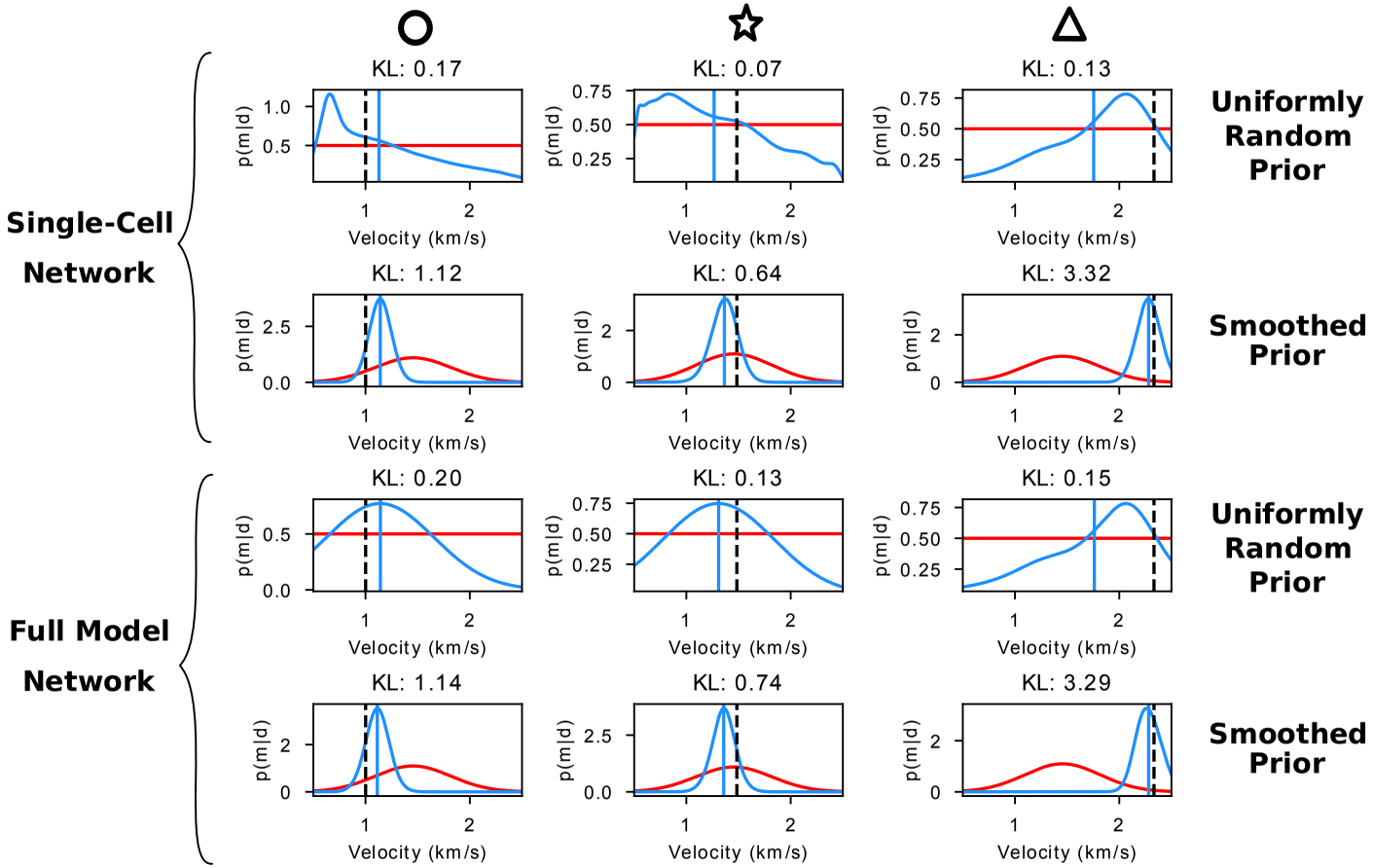

Figure 5.6 Posterior pdfs (blue curves) compared to the prior pdfs (red curves) for the 16 x 16 grid models for three locations shown in the top-right model of Figure 5.4: circle (left), star (middle), triangle (right). The rows show results from: (row 1) Separate-cell MDN's using Uniformly random training dataset. (row 2) Separate-cell MDN's using the smoothed training dataset. (row 3) Full-model MDN using Uniformly random training dataset. (row 4) Full-model MDN using the smoothed training dataset. The mean of the posterior is shown by the blue solid line and the true velocity value by a black dashed line. Corresponding KL divergence values are shown above each result.

5.3.3 Model Resolution

Our networks are trained on two sizes of grid cell, a coarser 8 x 8 grid and a finer 16 x 16 grid. Figures 5.4 and 5.5 show the results for varying grid size. Training on the finer grid induces a factor of 4 more parameters to estimate from the same data. This means that a larger training set size would be needed to sample the increase in image dimensionality. It would be impossible to sample densely the 256-dimensional space spanned by a 16 x 16 grid, but as the examples show, the networks are still able to invert for some basic structural information (Figure 5.4c). When I train the networks with a stronger prior pdf I reduce the effective dimensionality of the problem by introducing a relationship between neighbouring pixels: essentially all prior models and hence most posterior models lie on a significantly lower dimensional manifold that is embedded with the 64- or 256- dimensional spaces. In that case I can obtain reasonable estimates of the true velocity models regardless of grid size (Figure 5.5c).

5.3.4 Type of network

For each of the four training sets I trained networks in two different ways. First I trained separate networks to estimate marginal pdf's in each cell so that each network has fewer parameters (α_{ij}, μ_{ij}, σ_{ij}) to estimate. Note that this does not reduce the dimensionality of the overall problem as each velocity cell in the model contributes to the travel time values, and the velocity in any cell depends on the cells surrounding it even if I do not directly invert for them within the same network. It is important to remember that in this case I do not obtain explicit information about trade-offs between neighbouring cells. Those trade-offs are already integrated into the marginal pdf's in equation 2.2.

I also trained networks to invert for slowness in every cell of the model at once. This increases the number of parameters that the network must estimate but as a result the trade-off between velocity values in adjacent cells can be explored. Examples of the joint marginal pdfs from the central model in Figure 5.4a are shown in Figure 5.7: the 2D pdfs show few signs of non-linearity, and virtually no indication of the trade-offs that one would expect between velocities

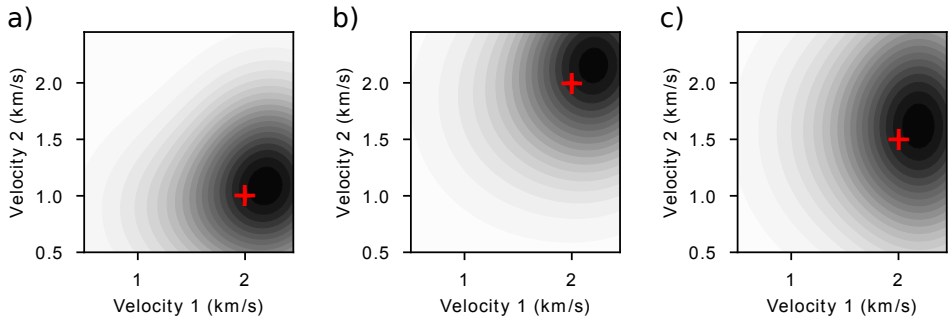

Figure 5.7 Joint pdfs comparing the a pixel inside the velocity high of the central model in Figure 5.4a. Velocity 1 is the velocity of a cell in the centre of the velocity high. Velocity 2 is the velocity of a cell a) in the background velocity, (b) at the centre of the velocity high (not the same cell as Velocity 1) and (c) at the edge of the velocity anomaly.

in neighbouring cells. This indicates that the results of these networks are unlikely to provide reliable uncertainties.

For models on a coarser grid (Figures 5.4 and 5.5 rows b and d), networks perform similarly when using the single cell networks or the full model networks. For models trained on a finer grid, the full model networks perform significantly better than the single cell network as shown in Figure 5.4. This is almost certainly because the dimensionality of the problem when training single-cell networks is too large, but by giving the network information about the velocities in neighbouring cells it can better resolve the velocities. This difference is less noticeable when using stronger prior information (Figure 5.5b and d).

5.3.5 Uncertainty Loops

A key problem in the field of nonlinear inversion is that there are no standard solutions to which estimated posterior pdf's can be compared in order to verify their quality. In almost all papers that use synthetic tests to assess competing methodologies in high-dimensional problems, the main criterion applied is whether the mean or maximum-likelihood model fits the real (true synthetic) model that was used to generate the synthetic data. This provides no test at all on the rest of the pdf and indeed there is no reason why the mean *should* match the true model in unknown problems - the mean may even be a zero-probability solution (one

precluded by the data) (Tarantola, 2005). The maximum likelihood (or maximum posterior probability) model is an alternative, but usually an extremely volatile statistic of pdf solutions since those solutions are necessarily formed by focusing across the whole pdf rather than simply on its modes. We therefore require some independent property of posterior pdfs, the existence of which we can use to assess their veracity.

Loops or halos of high uncertainty have been shown to exist in solutions to all travel time tomography problems around anomalies with a spatially sharp and strong contrast in velocity compared to their surroundings (Galetti et al., 2015). Uncertainty loops exist due to non-linear aspects of wave physics and represent uncertainty in the *shape* of such anomalies. They are observed most clearly in fully non-linearized tomographic inversion problems in which rays, velocities and travel times are all varied in concert for each sample considered. We can therefore use the existence of loops in posterior uncertainty as a criterion to check their quality in models with strong and spatially sharp contrasts.

Figure 5.8 shows the standard deviations (bottom row) for the results of networks trained on an 8 x 8 grid. Only the networks trained using the training set of Uniformly random velocities (Figure 5.8f and h) exhibit signs of an uncertainty loop. I include the mean (middle row) for comparison of the shape of the velocity anomaly to the loop that surrounds it. The difference between the two priors is clear when comparing Figures 5.8f, 5.8g and 5.8h: for a smoothed prior (Figure 5.8g) the maximum uncertainty is predicted to be in the centre of the anomaly as opposed to the other two images where the uncertainty is lowest at the centre of the anomaly and highest on the margins as expected. However, when inverting for the full-model in a single network (Figure 5.8h) the loop is not as well defined as in Figure 5.8f. Together with the lack of clear trade-off relations in Figure 5.7 this is evidence that the full-model inversions are less robust than single-cell inversions: as the networks invert for many more parameters at once, they appear not to have been trained so as to fully represent the correct physics of the tomography problem.

The separate-cell networks (one network trained for each cell in the velocity model) allow us to estimate the full marginal posterior probabilities for all cells in the model, and these posterior distributions show how the network represents uncertainty. I show the pdfs for 3 points in the model: inside the velocity anomaly (star), at the edge of the anomaly (triangle), and in the background velocity (circle), where the locations are shown in Figure 5.8a. I can see for the 8 x 8 model using the Uniformly random training set (Figure 5.9a and c) the posterior pdf at the edge of the anomaly has a larger uncertainty indicating that the range of possible velocities spans the velocity of the anomaly and that of the background velocity. This is expected at the edge of an anomaly, the boundaries of which are uncertain: the cells could either be inside or outside of the anomaly, and could therefore assume values of the anomaly (low velocity) or the background model (high velocity). This is the maximum range of velocities expected across the model, hence the largest uncertainties should be around anomaly edges (Galetti et al., 2015).

We do not see uncertainty loops in any model trained on the smoothed models. This makes sense because by imposing prior information that the model is relatively smooth we have removed the possibility to include the effect of spatially sharp contrasts between anomalies and the background velocity model, precluding the types of physical trade-offs that create uncertainty loops. This is represented in the pdfs (Figure 5.9b and d) where the uncertainty is much smaller than in (a) and (e) and where there is no noticeable increase in uncertainties at the boundary of the anomaly. Note that there is again a larger information gain for the results from the smooth training set as shown by the KL divergence values.

5.3.6 Realistic Velocity Models

Figures 5.10 and 5.11 show the results when applying the trained networks to other types of structures that might be encountered in geophysical or non-destructive testing applications. Figure 5.10 shows results using Uniformly random training set, whereas Figure 5.11 shows the equivalent results obtained using the smoothed training set. The models inverted on a coarser grid produce reasonable estimates

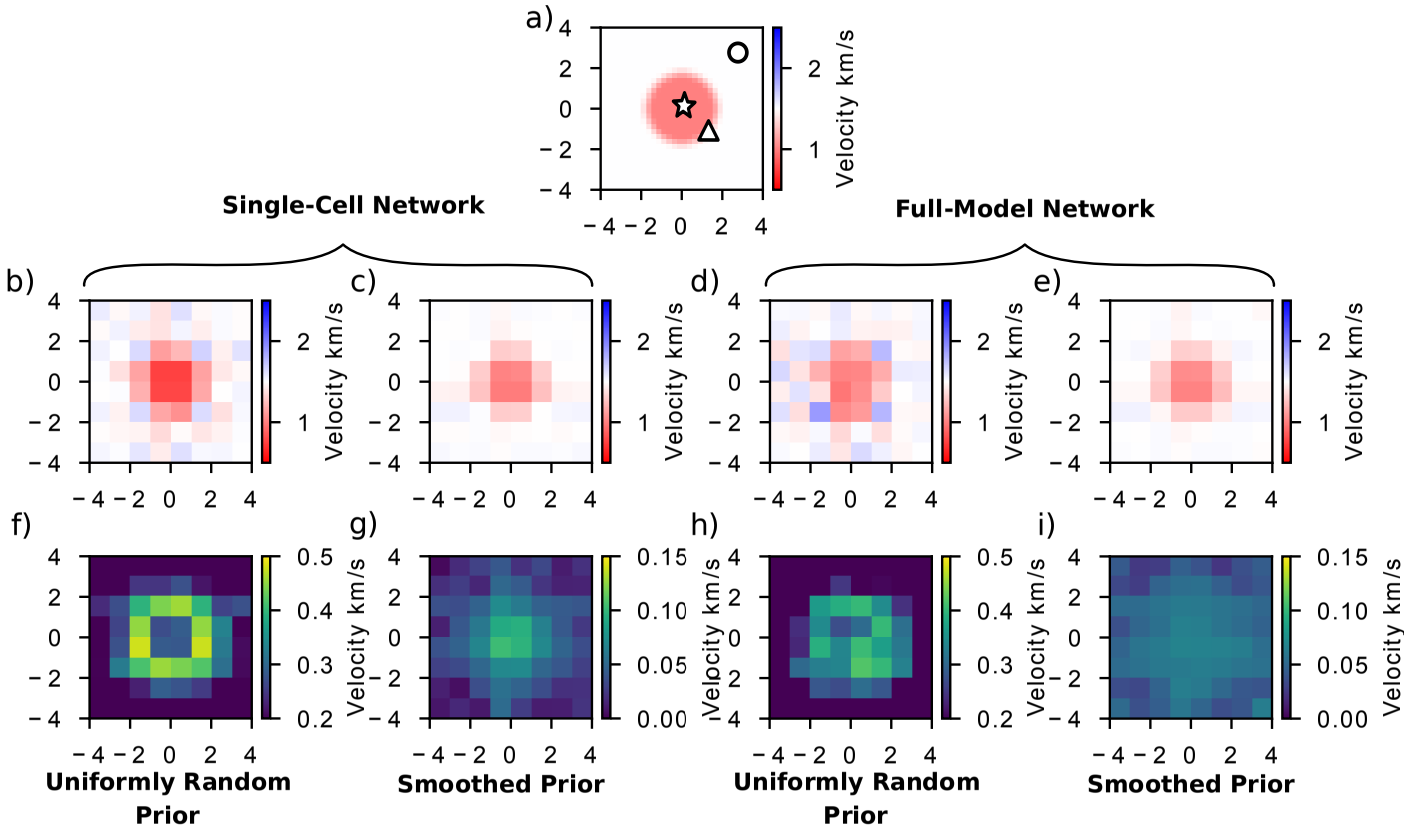

Figure 5.8 (a) True velocity model. For a separate-cell MDN, using a training set from a Uniformly random distribution, results shown are (b) mean velocities and (f) corresponding standard deviations. Using the same type of network with a training set of spatially smoothed velocities I obtain (c) mean velocities and (g) standard deviations. For a full-model MDN, using a training set from a Uniformly random distribution I obtain (d) mean velocities and (h) standard deviations. Using the same type of network with a training set of smoothed velocities I obtain e) mean velocities and (i) standard deviations.

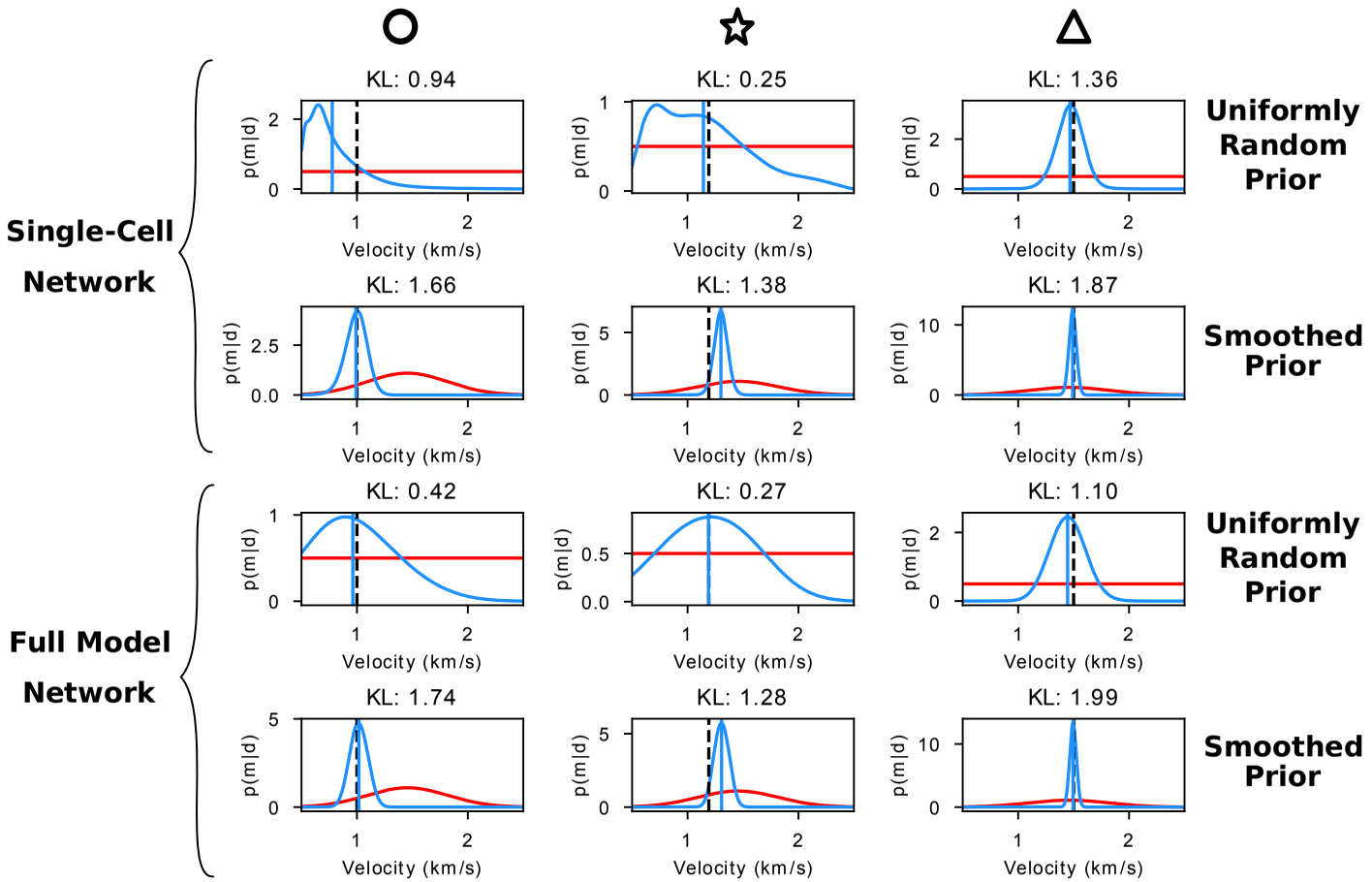

Figure 5.9 Posterior pdfs (blue curves) compared to the prior pdfs (red curves) for the 16x16 grid models for three locations shown in the true model of Figure 5.8: circle (left), star (middle), triangle (right). The rows show results from: (Row 1) Separate-cell MDN's using Uniformly random training dataset. (Row 2) Separate-cell MDN's using the smoothed training dataset. (Row 3) Full-model MDN using Uniformly random training dataset. (Row 4) Full-model MDN using the smoothed training dataset. The mean of the posterior is shown by the blue solid line and the true velocity value by a black dashed line. Corresponding KL divergence values are shown above each result.

of the velocity models using either prior pdf, however for the smoothed prior all the models, regardless of grid size, are recovered fairly well. Figure 5.12 shows the uncertainty maps for a coarse grid model trained using both types of prior information and inverted using the separate-cell MDN models. When inverting the models with a Uniformly random prior (Figure 5.12b) the uncertainty maps show a higher uncertainty at the anomaly interfaces (as expected by analogy with the uncertainity loops above), thus helping to define uncertainty in the model geometry, whilst the results from the smooth prior miss this extra infomation.

5.4 Discussion

I compared different methods of mixture density network inversions to estimate tomographic posterior probability density functions. When using datasets with little prior information (Figure 5.2a and 5.2b) the networks struggle to estimate more than the simplest of velocity models: due to the curse of dimensionality it is simply not possible to provide a sufficient density of prior samples on which to train the MDN. Including stronger prior information in the examples by training on smoothed velocity models (Figure 5.2d and 5.2e) improves inversion results, although the networks are no longer able to image sharp velocity contrasts, nor estimate uncertainty in the shapes and locations of spatially sharp velocity anomalies, as information about such models is not contained in the training set. The tests indicate that the prior pdf is the most important factor in improving a network performance since it restricts both the training set and inversion results to a more constrained (effectively lower-dimensional) manifold embedded within the high-dimensional parameter space. This manifold is more densely sampled than the full space thus improving network training and performance. All test models inverted using the stronger prior information give higher SSIM and KL divergence values compared to those using weaker priors, regardless of grid size or how many pixels were inverted with each network. Also, the two circular anomalies in Figure 5.5 are symmetrical and this symmetry is also shown in all of the smooth-prior inversion results which is not seen in the Uniform-prior results in Figure 5.4. Nevertheless, I show that when imposed prior information

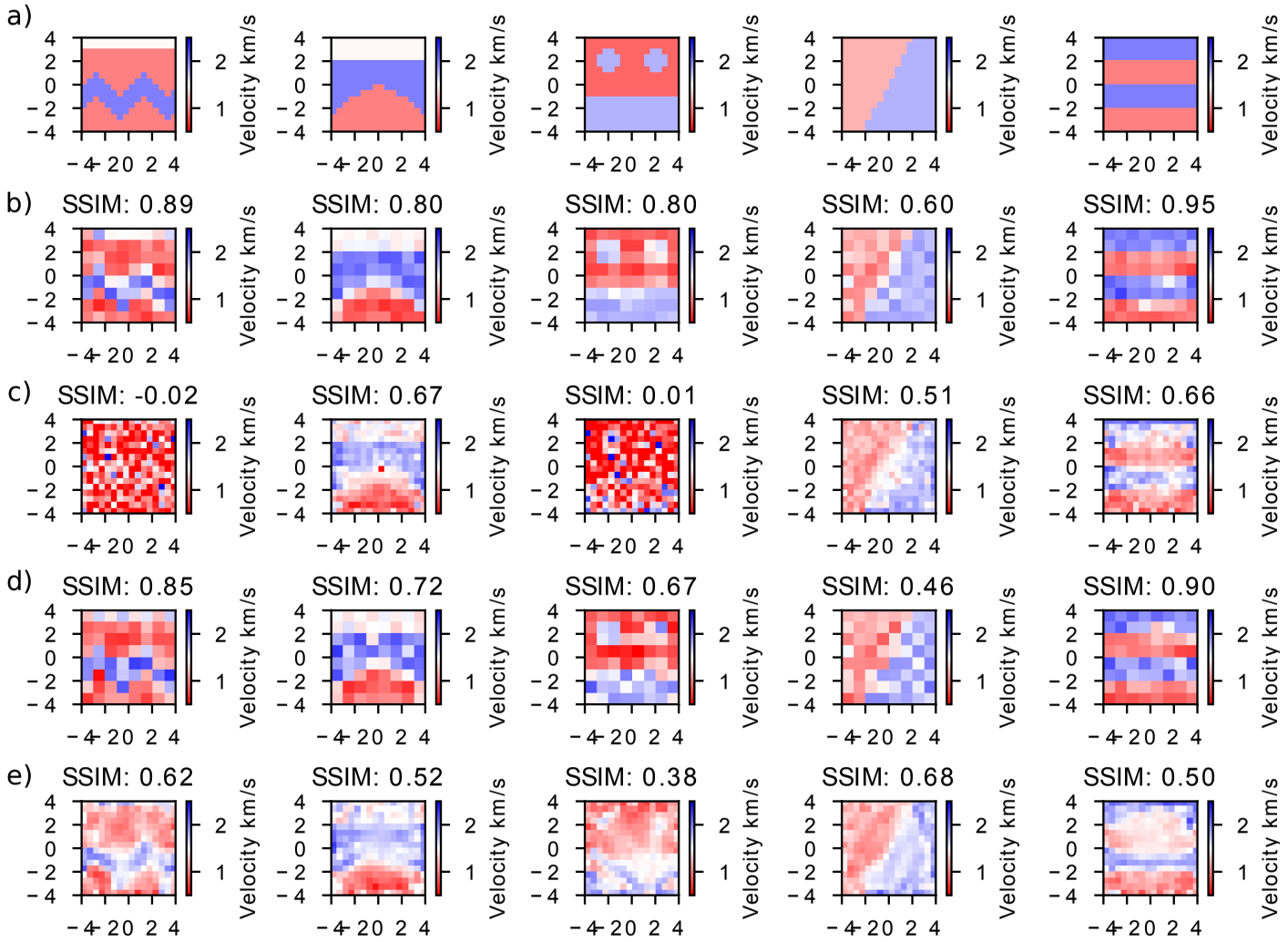

Figure 5.10 (a) True velocity models. Using a random generated training set from a Uniform distribution, mean velocities from separate-cell MDN inversions for (b) an 8 x 8 model and (c) a 16 x 16 model, and from full-model MDN inversion for (d) an 8 x 8 model and (e) a 16 x 16 model. The corresponding SSIM values are shown above each result (see Appendix D for definition of SSIM).

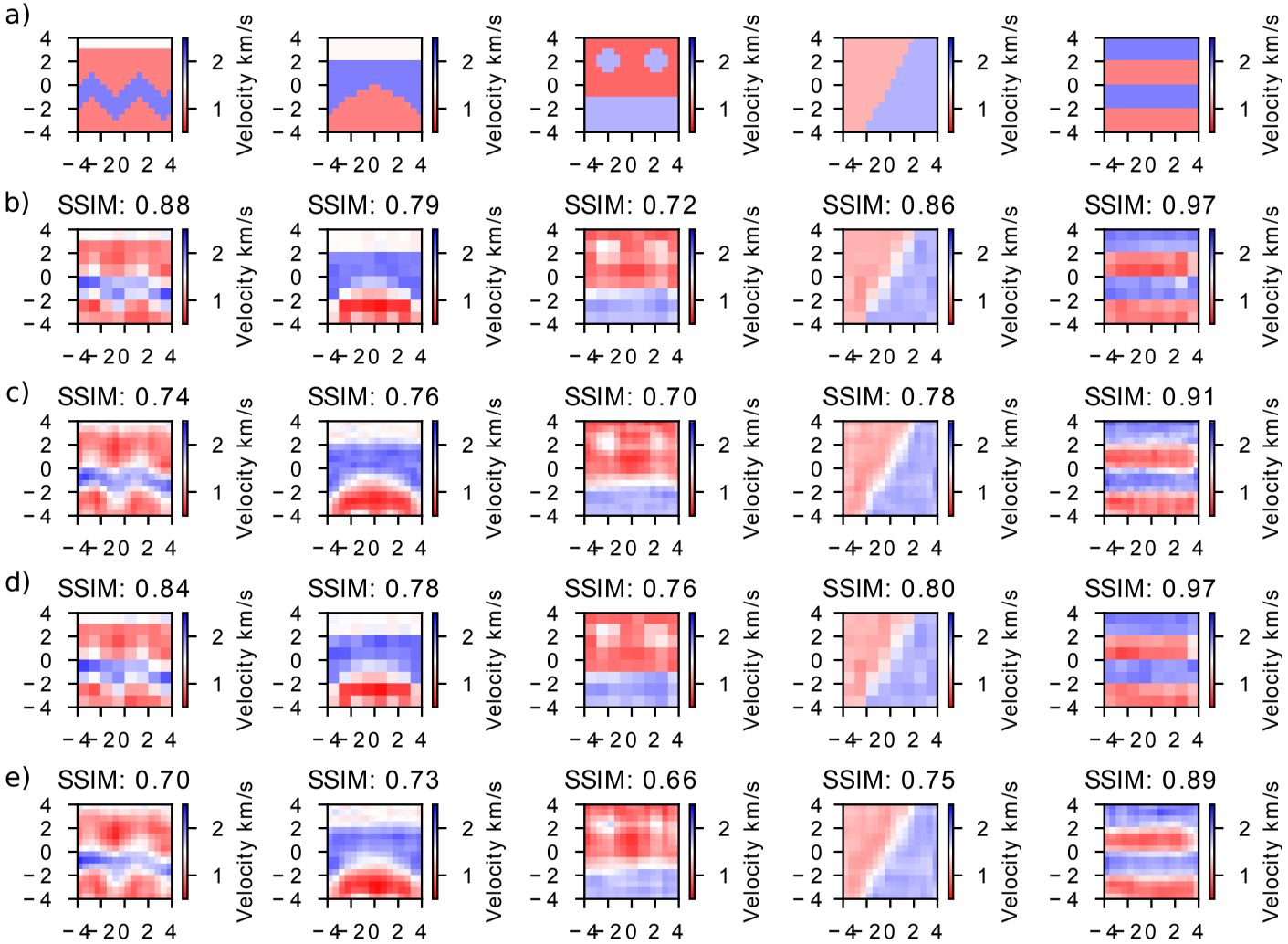

Figure 5.11 (a) True velocity models. Using a training set drawn of smoothed random models, mean velocities from separate-cell MDN inversions for (b) an 8 x 8 model and (c) a 16 x 16 model, and from full-model MDN inversions for (d) an 8 x 8 model and (e) a 16 x 16 model. The corresponding SSIM values are shown above each result (see Appendix D for definition of SSIM).

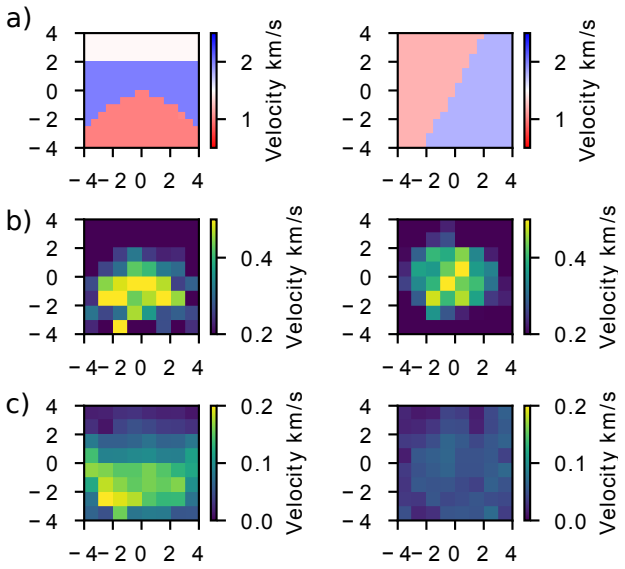

Figure 5.12 (a) True velocity models. For a separate-cell MDN, (b) the standard deviations using a generated training set from a Uniformly random distribution. Using the same network with a training set of smoothed velocities I obtain standard deviations (c).

is false (if the true model is rough but the prior precludes such models) then uncertainty results will be compromised as in Figure 5.8g and 5.8i. In other words, a clearly advantageous strategy for the future of neural network tomography is to invest effort in finding and using more sophisticated, and correct prior information (Curtis and Wood, 2004). Recent efforts in this direction include Walker and Curtis (2014a) who use expert elicitation to constrain prior multi-point geostatistics, Mosser et al. (2018) who use neural networks to parametrise geological prior information, and Nawaz and Curtis (2017, 2018, 2019) who use Markovian models and variational methods with embedded neural and mixture density networks to combine geological and geophysical information; these various directions appear to be strategically important for the future of this field.

I illustrate the differences in the KL divergence values in Figure 5.13. The top graph shows histograms of KL values obtained when networks are applied to all synthetic test data for the four different prior and network training types for the 8 x 8 grid model, and the bottom graph is similar but for 16 x 16 models. Both plots confirm that training with a stronger prior increases the information gain in the posterior as was indicated in Figure 5.6. Notice that this is not necessarily

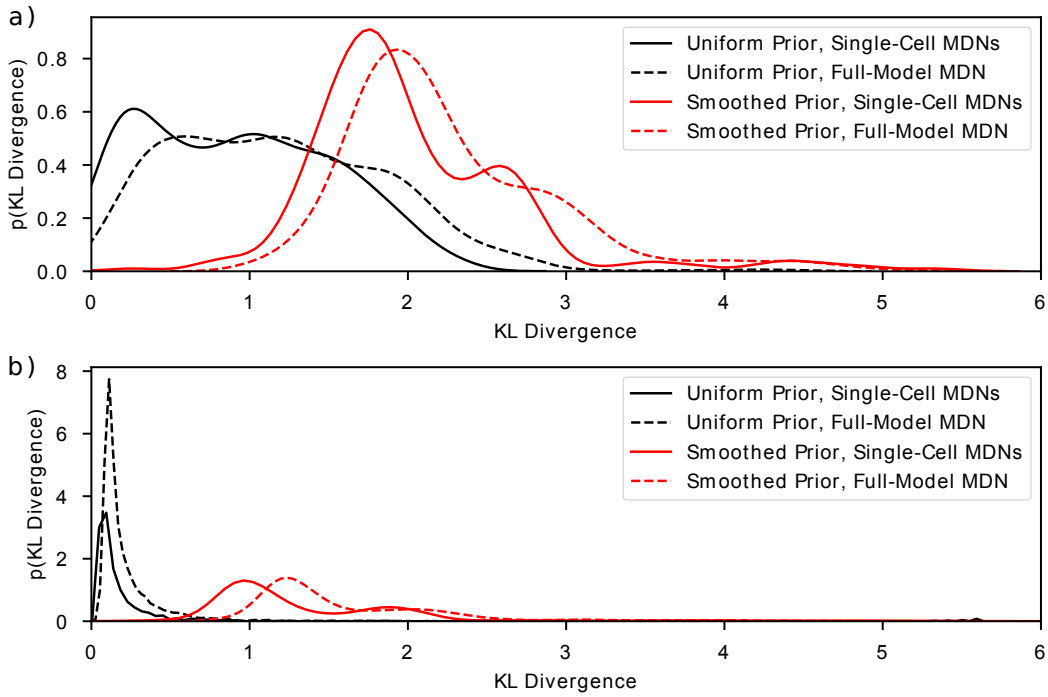

Figure 5.13 Histograms of KL divergence values for results of inverting synthetic data for all models in the test set. a) 8 x 8 models, b) 16 x 16 models.

an intuitively obvious result: if prior information is weaker or less informative, I might expect the data to add relatively more information, compared to the case where prior information is stronger. I therefore suspect that this result indicates that I simply can not train the MDN's in the case of weaker prior information and sparser training examples; even though by adding stronger prior information I should *decrease* the relative value of the data, this effect is out-weighed by the fact that I can better train the network and thereby extract *more* information from data.

The effect of increasing the number of cells in the model is also clearly highlighted: Figure 5.13a has higher KL values than Figure 5.13b. Interestingly, both plots show that training using a full-model inversion slightly increases the KL divergence, implying that the networks are making use of the relationship between adjacent pixels to better constrain the posterior pdfs.

5.4.1 Inference limits

When creating the training dataset I set hard bounds on the grid cell velocities, thus limiting the range of velocity models that should be found using the trained networks. Figure 5.14 shows the inversion of a model at the limits of all training datasets. The middle row shows results when using the Uniformly random training dataset: none of the inversions give reliable results. Although the network trained to invert the full model at once performs slightly better, all networks produce extremely poor results. This is expected as the velocity model has a background velocity at the lower limit of the training sets, $0.5km/s$ and an anomaly at the upper limit, $2.5km/s$. This is an extreme example that is not likely to have proximal samples in the training set, therefore the results are expected to be poor.

The same model lies outwith the dataset with a stronger prior as well, but networks appear to recognise that there is a velocity anomaly. However, since the prior dataset used is smoothed, strong contrasts are precluded and none of the networks give accurate velocity information, despite being able to represent the geometry of the structure.

5.4.2 Inversion Speed

As this is a prior sampling method the training dataset must be created in advance. It took $t_{prior} = 11$ hours, to create the training dataset of 2.5 million samples using 5 CPUs. However, this only needs to be done once; even if more prior information becomes available I may be able update the prior using the *prior replacement* method of Walker and Curtis (2014b) or the resampling method of Sambridge (1999) rather than calculating entirely new training examples.

In this work each network took between 1-2 hours to train (converge). For the 8 x 8 grid models with an ensemble of 8 networks when training the network for each grid cell separately, I required $8 \times 8 \times 8 = 512$ networks in total and for the 16 x 16 models with an ensemble of 4 networks I required $16 \times 16 \times 4 = 2048$ networks. However, the training of each network is independent of others so the process can easily be parallelised and using 50 cores a full training run for the larger 16 x 16

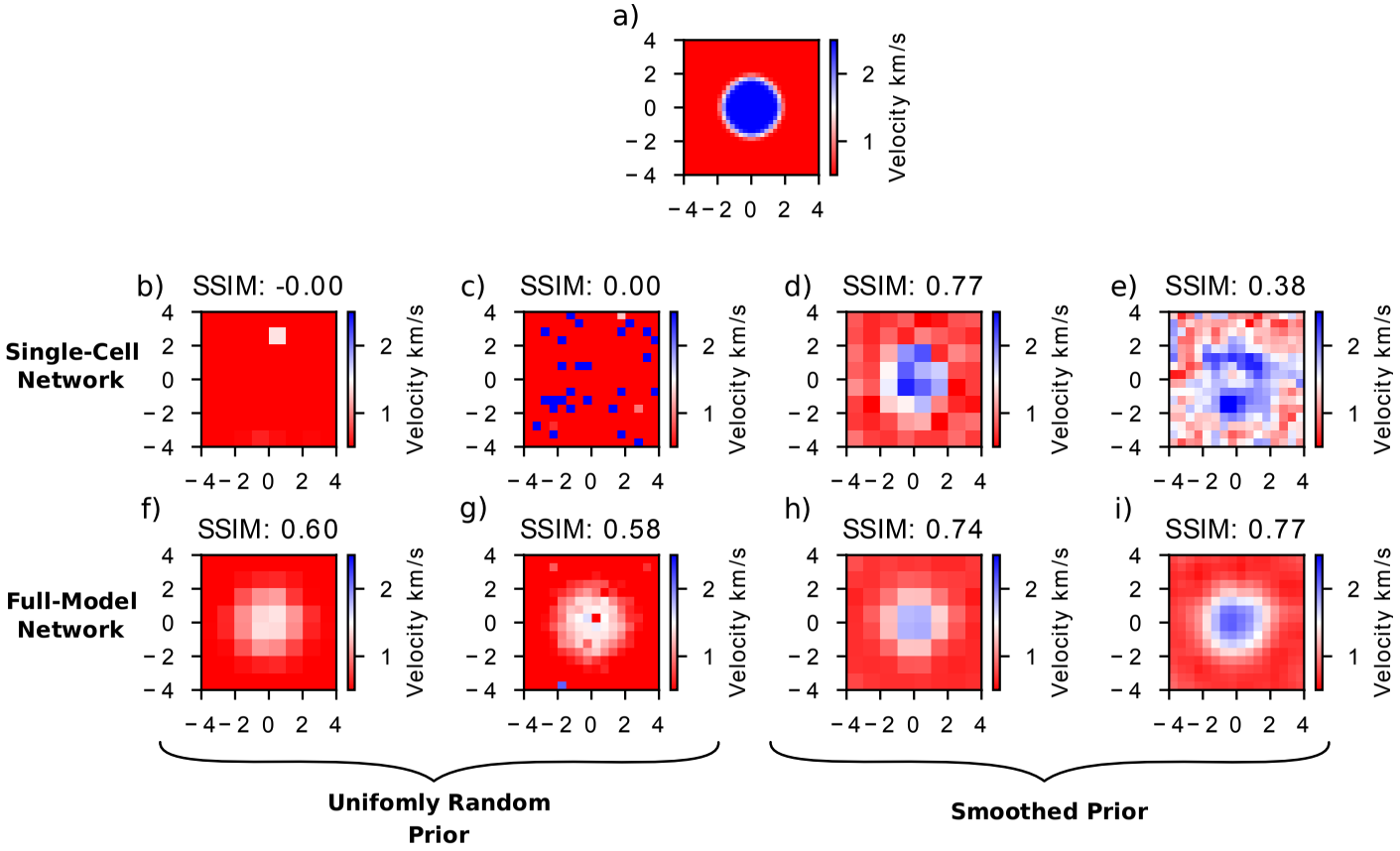

Figure 5.14 (a) True velocity model. For a separate-cell MDN, using a training set from a Uniformly random distribution, results shown are mean velocities for (b) an 8 x 8 model and (c) a 16 x 16 model. Using the same network with a training set of spatially smoothed velocities I obtain mean velocities for (d) an 8 x 8 model and (e) a 16 x 16 model. For a full-model MDN, using a training set from a Uniformly random distribution I obtain mean velocities for an (f) 8 x 8 model and (g) a 16 x 16 model. Using the same network with a training set of spatially smoothed velocities I obtain mean velocities for an (h) 8 x 8 model and (i) a 16 x 16 model. Corresponding SSIM is shown above each result. The colour axis has been clipped to the velocity bounds of the training set ($0.5km/s, 2.5km/s$).

grid model takes $t_{train} = 80$ hours of real clock time. For the full-model networks only one network is trained for all cells so the total training time is much lower: each network takes around 3 hours to train so training 10 networks only takes 30 hours without running them in parallel. This process could be reduced to 3 hours by using only 10 cores and reduced further by training each network across cores. The advantage of an MDN is the speed of inversion after training: once a network is trained new inversions take a fraction of a second, even on a standard desktop computer. Computational efficiency is therefore gained only when the trained networks will be applied to many different data sets.

Monte Carlo methods are known to be computationally expensive (Bodin and Sambridge, 2009) and a fully non-linear Markov chain Monte Carlo (McMC) tomographic inversion can take weeks or months of compute time. Monte Carlo methods use posterior sampling so for every new inversion a new sample set must be performed. This is often a far less demanding sampling task than sampling with similar density of samples from the prior since high probability parts of the posterior pdf usually span a significantly smaller volume of parameter space. Nevertheless, neural network methods are advantageous over traditional Monte Carlo methods when n repeated inversions of similar data types are to be performed provided that $n > \frac{(t_{prior}+t_{train})}{t_{MC}}$, as the computationally expensive sampling step only needs to be performed once and the network-based inversion becomes faster. In a tomographic setting this could be useful for monitoring purposes, where data collected periodically from the same set of sources and receivers can be inverted with the same network(s) each time new data arrives.

5.4.3 Training Flexibility

In this work I train networks assuming that the data (travel times) are recorded with exactly the same data acquisition geometry as was used for training. It would also be possible to train more flexible networks that account for missing data. For example, one could augment the training set with additional samples constructed from the same data-model pairs $(\mathbf{d}_i, \mathbf{m}_i) : i = 1, ..., N$ but with a certain number of travel time values in the dataset randomly set to 0, to indicate a missing value

(De Wit et al., 2013). Then new datasets with missing values (for example due to noisy stations causing errors in travel times) can be inverted by the same network.

Data from new receivers added after training the network will not be able to be use. However I can create a new training set containing only the data from the added receiver station and fine tune the original network by using the original network parameter values as a starting point for training optimisation. This has the advantage that the training process will be much faster.

5.5 Conclusion

I present neural network-based, non-linear inversion methods applied to a 2D travel time tomography problem to estimate posterior probability density functions. The flexibility of mixture density networks mean that I can provide uncertainty estimates for 2D velocity maps. I show that the prior information used to create the training dataset is the most important factor in providing accurate velocity estimates and uncertainties as such information effectively reduces the dimensionality of the tomography problem. However, as with all Bayesian inversions if we impose false prior information we can lose important information about uncertainties. By training networks to invert for a full tomographic model at once, we can also understand the relationship between velocities in neighbouring pixels; however the number of parameters in the inversion increases substantially, and training for accurate models proves to be significantly more difficult. I compare the speed of neural network inversion to more standard Monte Carlo methods and determine that for many repeated inversions such as occur in monitoring situations, MDNs may out-perform Monte Carlo methods in terms of computational cost.

Chapter 6

Estimating Subsurface Density by Full Waveform Inversion of Acoustic Reflections using Interferometric and Marchenko Methods

Seismic data are used to image subsurface structures and to estimate material parameters at scales ranging across global tomography, upper crustal resource exploration and monitoring, and shallow engineering applications. Density ρ is an important material parameter that is difficult to estimate using seismic data due to its inherent trade-off with velocity v in the acoustic impedance contrast $z = \rho v$ which is the parameter that primarily controls reflection amplitudes. As a consequence, density is usually inferred from Gardner's equation which relates density to P-wave velocity using a semi-empirical formula or by amplitude variation with offset analysis. We propose and test a new method to estimate density in the subsurface using an acoustic approximation, and a formulation of seismic interferometry that contains a linear dependency on density alone. We demonstrate the method on synthetic data that represents recorded wavefields at subsurface receivers in a borehole, and on wavefields from subsurface virtual receivers created by Marchenko redatuming away from a borehole. The method returns reasonable density estimates at the borehole location, but practical limitations of the Marchenko method render the results less accurate where no borehole

is present. The key factor that deteriorates the density result in the latter case is the absolute scaling of Marchenko-derived Green's function estimates. We present a new way to normalize the Marchenko method that solves the problem in layered media. Through the methodology proposed herein, future research that improves estimates of Marchenko amplitudes in more heterogeneous media will contribute directly to our knowledge of subsurface density.

6.1 Introduction

Seismic data are used to image subsurface structures and material properties over length scales ranging across global tomography (Woodhouse and Dziewonski, 1984; Su et al., 1994) and reflection imaging (van der Hilst et al., 2007), crustal scale resource exploration (Claerbout, 1985), subsurface storage monitoring (Lumley, 2010) and bedrock depth estimation for civil engineering (Hunter et al., 1984). One important material property is density ρ. This property is necessary in order to interpret recorded seismic wave amplitudes (for example in waveform inversion studies) because acoustic or elastic impedance $z = \rho v$, where v is seismic velocity, controls amplitudes of reflected waves (Connolly, 1999). Density is also important in geological or petrophysical interpretation because it is closely related to rock porosity and fluid properties, and can be a useful discriminator between geological facies.

Despite its importance, subsurface density estimates are generally very uncertain due both to the trade-off of density with seismic velocity given measurements of impedance contrast between strata, and to the inherent non-uniqueness in density estimation from potential-field data such as gravity measurements (Blakely, 1995). When inverting for density using wave equation based inversion schemes, such as Full Waveform Inversion (FWI), trade-offs between P- or S-wave velocity and density deteriorate results (Jeong et al., 2012): density is often estimated using Gardner's equation which relates P-wave velocity to bulk density (Gardner et al., 1974), but this can bias results by overlooking anomalies in density that are not coupled to velocity variations. Prieux et al. (2013) use multi-parameter visco-acoustic FWI to jointly invert for density, P-wave velocity and quality fac-

tor. They perform a first inversion step to reconstruct the dominant parameter, P-wave velocity, before performing a joint inversion for all three parameters. This produces a more stable update to the initial smooth density model compared with a single step joint inversion. However, changing the inversion protocol cannot, on its own, remove implicit trade-offs between density and velocity in the physics of the problem. On a global scale, Blom et al. (2017) show that density can be estimated using waveform inversion despite also showing that density variations have a relatively small effect on the seismic wavefield compared to the effects of velocity variations. Density can be retrieved from AVO inversion (Dębski and Tarantola, 1995; Downton and Lines, 2004), however this can be difficult without long offsets (Li, 2005).

In this chapter I exploit new opportunities to constrain density, offered by seismic interferometry and Marchenko theory. Seismic interferometry estimates the Green's function (wavefield that would be recorded) between the locations of two receivers, simulating the situation where one of the receivers acts as a virtual (imagined) impulsive source. This is achieved by integrating the crosscorrelation, convolution or deconvolution of wavefields from a boundary of other, real sources (Wapenaar, 2004; van Manen et al., 2005, 2006; Wapenaar and Fokkema, 2006; Wapenaar et al., 2011). Seismic interferometry has been used in seismic exploration for imaging under complex overburdens (Bakulin and Calvert, 2006; van der Neut et al., 2011), interferometric velocity analysis (King and Curtis, 2011; King et al., 2011), ground roll removal (Curtis et al., 2006; Halliday et al., 2007, 2010; Duguid et al., 2011) and salt-flank imaging (Xiao et al., 2006).

By contrast, Marchenko methods use surface seismic reflection data and a macro-velocity model to estimate the Green's functions from a virtual source inside a medium to receivers either in the subsurface or on the surface, without requiring a receiver at the virtual source location. They thus directly redatum surface seismic sources to subsurface locations. Methods have been formulated for acoustic (Broggini et al., 2012; Wapenaar et al., 2014) and elastic (da Costa Filho et al., 2014; Wapenaar, 2014) media without a free surface, and in acoustic media for the case that free surface multiples are included (Singh et al., 2015; Ravasi, 2017; Slob and Wapenaar, 2017). The method can then be used directly for

subsurface seismic imaging (Wapenaar, 2014; da Costa Filho et al., 2015; Ravasi et al., 2016; da Costa Filho et al., 2017; Singh and Snieder, 2017), for imaging combining surface reflection data and horizontal borehole data (Liu et al., 2016), or for finding travel-time changes in time-lapse data near horizontal boreholes (Liu et al., 2017). A set of methods called adaptive Marchenko redatuming have been shown to be robust in field data examples (van der Neut et al., 2015; Staring et al., 2017).

Meles et al. (2015) showed how a combination of Marchenko redatuming and seismic interferometry could be used to predict internal multiples - waves that reflect more than once from interfaces in the subsurface. Using up and downgoing wavefields from Marchenko methods they reconstruct the internal multiple wavefield using convolutional interferometry. The multiples can then be adaptively subtracted from the original seismic data to reveal the primary (singly-scattered) reflections. Meles et al. (2016) used a similar method to construct primary reflections directly. van der Neut et al. (2018) derive an alternative Marchenko equation that uses a two-way traveltime surface of a horizon instead of a macro-velocity model. The resulting data set can be used for internal multiple removal by adaptive subtraction without using convolutional interferometry as in the method of Meles et al. (2015).

Since both the multiples and the primaries can be constructed, the methods of Meles et al. (2015, 2016) can be used to predict the complete surface seismic data from the data itself through an interim step of seismic interferometry. While this may seem circular, it is important because, in acoustic media, convolutional interferometric theory exhibits a dependence on density alone, rather than on the product of density and velocity in the impedance. This suggests that it may be possible to invert the theory of Meles et al. (2015, 2016) to estimate density without the usual trade off with velocity.

I investigate this method to estimate density in the subsurface under the acoustic approximation. The method is demonstrated on synthetic data that simulate both recorded wavefields at subsurface receivers in a borehole, and wavefields from virtual receivers created by Marchenko redatuming. Finally

remaining issues are discussed that must be overcome in the case where Marchenko data are used, and hence the challenges for applications to real data.

6.2 Theory

6.2.1 Convolutional Interferometry

Seismic interferometry is a technique to estimate the Green's function between a pair of receivers, that is, the seismograph that would have been observed if one of the receivers had been an impulsive source that was recorded by the other. This is achieved by integrating the crosscorrelation, convolution or deconvolution of wavefields from a boundary of other, real sources (Wapenaar, 2004; van Manen et al., 2005; Curtis et al., 2006; van Manen et al., 2006; Wapenaar and Fokkema, 2006; Curtis and Halliday, 2010). To summarize key aspects of the theory for our purposes I start with the reciprocity theorem of the convolution type. This relates two independent acoustic wave states that exist in the same medium, where a state is defined by a combination of material parameters, source distributions, boundary conditions and initial conditions (Fokkema and van den Berg, 1993; van Manen et al., 2005; Wapenaar and Fokkema, 2006)

$$\int_{\mathbb{D}} \left(\hat{p}_A \hat{q}_B + \hat{v}_{i,A} \hat{f}_{i,B} - \hat{q}_A \hat{p}_B + \hat{f}_{i,A} \hat{v}_{i,B} \right) d^3x = \oint_{\partial\mathbb{D}} \left(\hat{p}_A \hat{v}_{i,B} - \hat{v}_{i,A} \hat{p}_B \right) n_i d^2x \quad (6.1)$$

where $\hat{p}$ is acoustic pressure, $\hat{v}_i$ is particle velocity, $\hat{f}_i$ is a force density source and $\hat{q}$ is a volume injection rate density source, subscripts A and B represent two different wavefield states occurring in a portion of the same medium $\mathbb{D}$ which is defined to be the interior of a boundary $\partial\mathbb{D}$, and where n is the outward-pointing unit normal vector to $\partial\mathbb{D}$. Consider a boundary $\partial\mathbb{D}$ that encloses one but not both sources, such as that in the schematic diagram in Figure 6.1. If we let the wavefields be pressure responses to an impulsive volume injection rate source and assume that external forces are zero, we can set the terms in Equation 6.1 to the quantities in Table 6.1. Note that for this case, $\hat{q}_B = 0$ because point x_1 is located outside of boundary $\partial\mathbb{D}$, and therefore $\delta(x - x_1) = 0$.

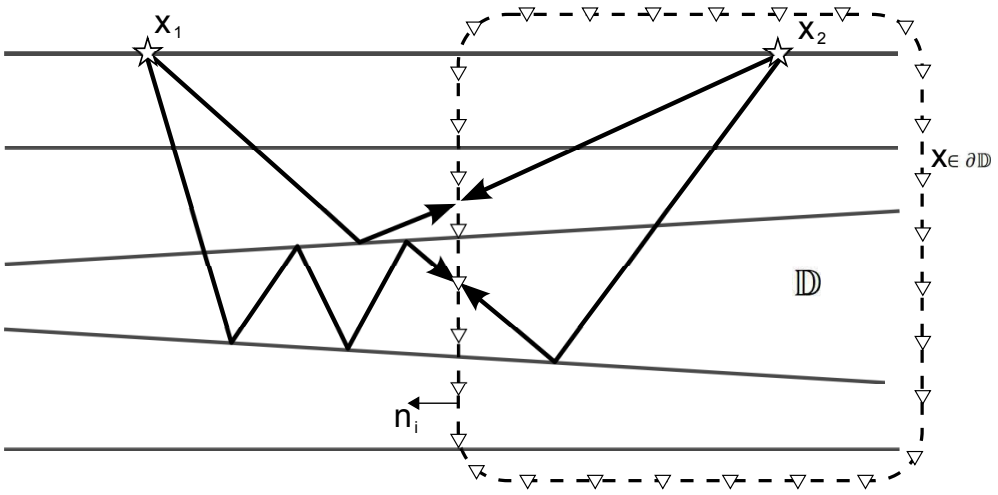

Figure 6.1 Schematic geometry for convolutional interferometry. Closed boundary $\partial\mathbb{D}$ with normal vector $\mathbf{n}$ is represented by the black dashed line and white triangles represent receivers. Sources are represented by stars and black arrows represent example ray paths of acoustic wavefields.

State A		State B	
$q_A(x,\omega)$	$\delta(x-x_2)$	$q_B(x,\omega)$	0
$p_A(x,\omega)$	$G(x,x_2,\omega)$	$p_B(x,\omega)$	$G(x,x_1,\omega)$
$v_{i,A}(x,\omega)$	$-(j\omega\rho(x))^{-1}\partial_i G(x,x_2,\omega)$	$v_{i,B}(x,\omega)$	$-(j\omega\rho(x))^{-1}\partial_i G(x,x_1,\omega)$
$f_{i,A}(x,\omega)$	0	$f_{i,B}(x,\omega)$	0

Table 6.1 Definitions of wavefield quantities in Equation 6.1 when the free surface is not present and the geometry of $\partial\mathbb{D}$ is defined in Figure 6.1.

Substituting the identities in Table 6.1 into Equation 6.1 and applying Gauss's Theorem gives (van Manen et al., 2005)

$$G(x_2,x_1,\omega) = \int_{x\in\partial\mathbb{D}} \frac{1}{j\omega\rho(x)} \left[G(x,x_2,\omega)\partial_i G(x,x_1,\omega) - \partial_i G(x,x_2,\omega)G(x,x_1,\omega)\right] n_i \partial\mathbb{D} \tag{6.2}$$

where $\rho(x)$ represents material mass density at the location x, the integration is over points x on boundary $\partial\mathbb{D}$, and x_1 and x_2 are interpreted as receiver locations. Derivative ∂_i is the *ith* component of the outward-pointing gradient at the boundary $\partial\mathbb{D}$. $G(x_2,x_1,\omega)$ represents the acoustic pressure Green's function in the angular frequency (ω) domain measured at x_2, from a volume injection rate source at x_1 in an inhomogeneous medium. The wavefield $\partial_i G(x,x_1,\omega)$ represents

the wavefield from a source at x_1 recorded by a so-called dipole receiver at x on the boundary $\partial\mathbb{D}$, where the latter is a receiver that measures the spatial derivative of the wavefield in direction n (Wapenaar and Fokkema, 2006).

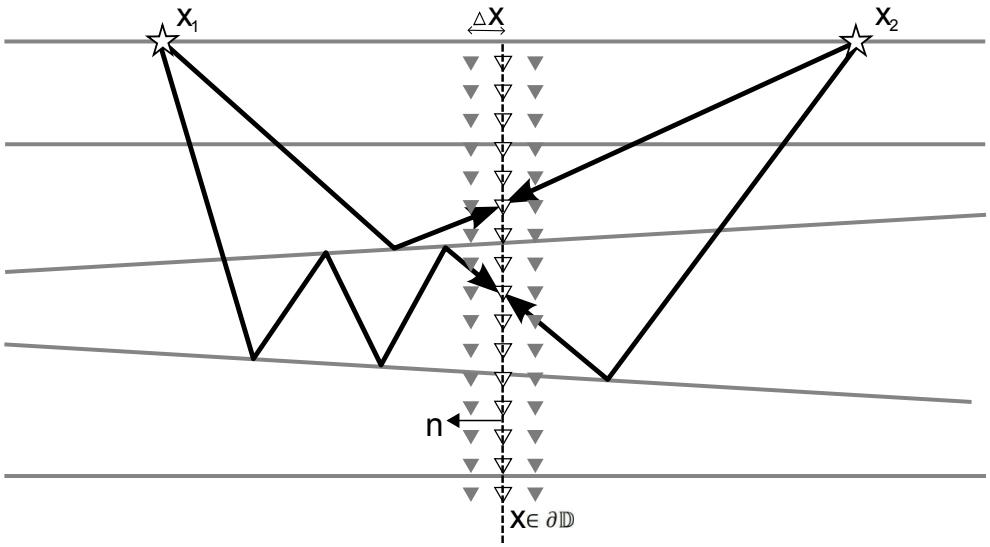

Figure 6.2 Schematic geometry for convolutional interferometry used herein to estimate density. The vertical boundary $\partial\mathbb{D}$ is represented by the black dashed line and receivers on the boundary are the central column of white triangles. The neighboring columns of grey triangles are the receiver locations used to calculate dipoles or derivative wavefields, and Δx is the horizontal distance between receiver columns. Sources are represented by stars and black arrows represent example ray paths of acoustic wavefields.

Equation 6.2 is the standard convolutional interferometry equation. The main contributions to the result $G(x_2, x_1, \omega)$ on the left come from regions of the boundary $\partial\mathbb{D}$ over which the integrand is stationary. These are regions around receivers on $\partial\mathbb{D}$ for which rays to x_1 and x_2 have similar orientation, and hence are receivers through which raypaths between x_1 and x_2 would pass (Halliday and Curtis, 2009). I assume that there are no strong heterogeneities that would allow waves exiting the boundary to the right to scatter back to location x, which implies that the right hand edge of $\partial\mathbb{D}$ can be ignored as those waves will never propagate between both locations x_1 and x_2. I further assume that the boundary $\partial\mathbb{D}$ extends to sufficient depth that wavefields crossing the lower portion of the boundary can be ignored. It should be noted that in reality this can be difficult to achieve because boreholes usually only extend down to depths of interest and not

below. No waves can reflect back from the top part of $\partial\mathbb{D}$ so this can be ignored. Thus it is sufficient to consider that the integral in equation 6.2 only spans a portion of the vertical boundary as shown in Figure 6.2 as this must include all remaining stationary phase regions and allow $G(x_2, x_1, \omega)$ to be constructed by Equation 6.2.

Equation 6.2 requires wavefield recordings at locations x in the subsurface. In the true 3D Earth the locations should normally span a 2D surface orientated perpendicularly to the page of Figure 6.2. Since one would rarely have access to such a subsurface grid of receivers, I use a 2D approximation to the theory and assume that the propagating waves between x_1 and x_2 remain in the plane of the page, and that the 2.5D geometrical spreading in recorded data has been converted to the equivalent 2D geometrical spreading using a suitable transform (Lomas and Curtis, 2018).

In practice, the recorded wavefields might be obtained in two ways: either using a vertical borehole containing physical down-hole receivers, or by a method that is able to estimate the wavefield response to surface sources at arbitrary points in the subsurface. If a borehole is used then the wavefields $G(x, x_1, \omega)$ and $G(x, x_2, \omega)$ can simply be measured in the borehole. The derivative wavefields can then be estimated from wavefields $G(x, x_1, \omega)$ and $G(x, x_2, \omega)$ (see Appendix E). It would be better if the derivative wavefields could be measured directly but currently there appears to be no standard technology capable of recording these fields in a borehole.

6.2.2 Marchenko Method

If the full wavefield response of a subsurface point is required without a borehole, such wavefields must be estimated by a method of downward continuation or extrapolation of fields measured at the surface. The Marchenko method has been developed in order to estimate the Green's functions from a virtual source (or virtual receiver) inside a medium using the single-sided reflection response of an inhomogeneous medium. This method is based on work by Rose (2002) who proposed a scheme to focus sound in a one-dimensional layered medium, using

an iterative sequence of operations that solve the so-called Marchenko equation. Broggini et al. (2012) showed that it is possible to retrieve the 1D Green's function of a virtual source in the Earth's subsurface at which location there was no physical source or receiver, using only the reflection response of the medium and an estimate of the direct wave from the virtual source (or to the virtual receiver). This method has since been extended to 3D acoustic media (Wapenaar et al., 2014) and elastic media (da Costa Filho et al., 2014; Wapenaar, 2014). Here I summarise the essential components of the Marchenko method, but a more complete explanation and derivation can be found in Wapenaar et al. (2014).

Key to the Marchenko scheme are focusing functions: these are wavefields that focus to a point at a certain depth, but only inside a reference medium which contains the true medium's reflectors above the focusing depth and is reflection free below the focusing depth. The focusing functions are therefore mathematical entities (rather than existing in the real Earth) that contain all true reflections from heterogeneity above the focusing depth, but after focusing they continue as a diverging down-going wavefield below the focusing point. They can be related to up- and down-going components of Green's functions, $G^{-}(x, x_0^{''}, t)$ and $G^{+}(x, x_0^{''}, t)$ respectively, decomposed at a subsurface recording point x given surface sources $x_0^{''}$ by

$$G^{-}(x, x_0^{''}, t) = \int_{\partial\mathbb{D}_0} \int_{-\infty}^{t} R(x_0^{''}, x_0, t-\tau) f^{+}(x_0, x, \tau) d\tau dx_0 - f^{-}(x_0^{''}, x, t) \quad (6.3)$$

$$G^{+}(x, x_0^{''}, t) = -\int_{\partial\mathbb{D}_0} \int_{-\infty}^{t} R(x_0^{''}, x_0, t-\tau) f^{-}(x_0, x, -\tau) d\tau dx_0 + f^{+}(x_0^{''}, x, -t) \quad (6.4)$$

where $f^{+}(x_0, x, \tau)$ is the inverse of the full transmission response from the surface to x, and $f^{-}(x_0, x, -\tau)$ is the corresponding reflection response (Slob et al., 2014). $R(x_0^{''}, x_0, t-\tau)$ is the surface reflection response which is assumed to have been measured and have had the free surface effects removed. The focusing functions $f^{+/-}$ focus at x and are calculated by the following iterative scheme:

$$f_k^{+}(x_0^{''}, x, t) = f_0^{+}(x_0^{''}, x, t) + \theta(t + t_d(x_0^{''}, x)) \int_{\partial\mathbb{D}_0} \int_{-\infty}^{\infty} R(x_0^{''}, x_0^{'}, \tau) f_{k-1}^{-}(x_0^{'}, x, t+\tau) d\tau dx_0^{'} \quad (6.5)$$

$$f_k^-(x_0^{''}, x, t) = \theta(t_d(x_0^{''}, x) - t) \int_{\partial\mathbb{D}_0} \int_{-\infty}^{\infty} R(x_0^{''}, x_0^{'}, t-\tau) f_k^+(x_0^{'}, x, \tau) d\tau dx_0^{'} \quad (6.6)$$

with

$$f_0^+(x_0^{''}, x, t) = T_d^{inv}(x, x_0^{''}, t) \quad (6.7)$$

Here θ is the Heaviside function which essentially acts as a window function, and T_d^{inv} is the direct arrival of the inverse of the transmission response which is usually approximated by the time reverse of the direct arrival and can be estimated from a smooth velocity model. After convergence the solutions for f_k^+ and f_k^- can be substituted into Equations 6.3 and 6.4 as f^+ and f^- to obtain Green's function estimates for $G^+(x, x_0^{''}, t)$ and $G^-(x, x_0^{''}, t)$. Using this method, responses from virtual sources in the subsurface to an array of surface receivers can be estimated using only two quantities: a spatially well sampled reflection response free from source and receiver ghosts and free surface multiples, and the direct transmission response which is usually estimated from a smooth velocity model. The resulting Green's function estimates can then be inserted into Equation 6.2.

6.2.3 Linear Inversion

Equation 6.2 shows that the Green's function $G(x_2, x_1, \omega)$ is linear in the reciprocal of density along the boundary $\partial\mathbb{D}$. Estimating $\frac{1}{\rho(x)}$ is therefore a linear inverse problem given all of the other terms in the equation. This can be written more concisely as

$$\mathbf{d} = \mathbf{Am}, \quad (6.8)$$

where $\mathbf{d}$ is the full recorded wavefield data between x_1 and x_2 or in other words the surface reflectivity (left side of Equation 6.2), $\mathbf{A}$ is the matrix containing the results of convolving wavefields $G(x, x_2, \omega) n_i \partial_i G(x, x_1, \omega) - n_i \partial_i G(x, x_2, \omega) G(x, x_1, \omega)$ in the integrand of Equation 6.2, and $\mathbf{m}$ is the vector of the reciprocal of density at locations along the boundary. We wish to use $\mathbf{d}$ and $\mathbf{A}$ to estimate $\mathbf{m}$.

To solve the problem I apply a least squares inversion scheme with Tikhonov regularization (Hansen, 1998). The objective function to be minimized is defined as:

$$\mathbf{J} = \min \left\{ \|\mathbf{Am} - \mathbf{d}\|_2^2 + \mu^2 \|\mathbf{Lm}\|_2^2 \right\} \quad (6.9)$$

where $\|\mathbf{Am}-\mathbf{d}\|_2^2$ is the data misfit measured using an l_2 norm, $\mathbf{L}$ is a second order difference operator, and μ is a constant known as the regularization or trade-off parameter. The solution that minimizes Equation 6.9 is given by

$$\hat{\mathbf{m}} = (\mathbf{A^T A} + \mu^2 \mathbf{L^T L})^{-1}(\mathbf{A^T d}) \tag{6.10}$$

which can subsequently be used to obtain an estimate for density because the model estimate $\hat{\mathbf{m}}$ consists of estimates of $\frac{1}{\rho(x)}$. The trade-off parameter μ defines the relative strength of constraints from the data versus the regularization, and herein is calculated using the L-curve method that plots a curve of the log of the norm $\|\mathbf{Lm}\|_2$ versus the log of the residual norm $\|\mathbf{Am}-\mathbf{d}\|_2$. The curve generally forms an L-shaped corner near the trade-off parameter value that is considered 'optimal' — the point where the errors of the two norms are equally minimized (Hansen, 1992).

6.3 Method

The theory explained in the previous section was used to create the algorithm shown in Box 1 to produce synthetic test data, and use them to estimate density in the subsurface model.

Equation 6.2 defines wavefields in terms of Green's functions, but if source signature deconvolution has not been applied then the reflectivity and borehole data will contain a band-limited source term. These will need to be included in Equation 6.2 and be balanced on both sides before application. For ease of calculation both sides are multiplied by $j\omega$ and the equation is converted to the time domain so that the left side of Equation 6.2 becomes the time derivative of the reflectivity. The expression then becomes

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\hat{R}(x_2,x_1,t)\otimes s(t)\\&=\int_{x\in\partial\mathbb{D}}\frac{1}{\rho(x)}\left[\hat{G}(x,x_2,t)\otimes\partial_i\hat{G}(x,x_1,t)-\partial_i\hat{G}(x,x_2,t)\otimes\hat{G}(x,x_1,t)\right]n_i\partial\mathbb{D}\end{aligned} \tag{6.11}$$

1. Define 3 equally spaced vertical boundaries very close together (Figure 6.2) and either record simulated wavefields from surface sources at all receiver locations down these simulated boreholes, or calculate the full wavefields that would have been recorded at the same locations using Marchenko methods.

2. Calculate the dipole responses $\partial_i G$ in Equation 6.12 by finite-difference methods using the two outer vertical boundaries of receivers (or using the approximation in Appendix E), then calculate $G(x, x_2)n_i\partial_i G(x, x_1) - n_i\partial_i G(x, x_2)G(x, x_1)$.

3. Use the L-curve method to estimate an 'optimal' value for μ, then apply Equation 6.10 to estimate $\hat{\mathbf{m}}$ and calculate density along the central vertical boundary.

4. If calculating wavefields with Marchenko methods, repeat steps 1-3 for vertical boundaries $\partial\mathbb{D}$ shifted laterally to each horizontal location at which density estimates are required.

Box 1 Algorithm to test the method to estimate density in the subsurface.

where $\hat{G}(t) = G(t) \otimes s(t)$ and $\hat{R}(x_2, x_1, t) = G(x_2, x_1, t) \otimes s(t)$ represents the measured reflectivity without deconvolution of the source signature $s(t)$ and $\otimes$ denotes convolution. For our tests the dipole or derivative wavefield is not measured directly but is approximated (in step 2 of the algorithm in Box 1) by the spatial derivative in the direction of the outward pointing normal to the boundary. This is calculated using two extra vertical lines of receivers either side of the central receiver (Figure 6.2) and the centered finite-difference scheme

$$\partial_i \hat{G}(x, x_1) = \frac{\hat{G}(x + \Delta x, x_1) - \hat{G}(x - \Delta x, x_1)}{2\Delta x} \tag{6.12}$$

where Δx is the distance between receiver pairs (Figure 6.2) and should be much smaller than the minimum wavelength of the model. In a real borehole no extra vertical lines of receivers would be available but given an estimate of medium slowness $u_r(x)$ and the direction of propagation θ of the wavefield at the borehole

receiver an approximation based on plane-wave approximation can be given by

$$\partial_i \hat{G}(x, x_1) = -u_r(r) \sin\theta \frac{\partial \hat{G}(x, x_1)}{\partial t} \tag{6.13}$$

a derivation of which can be found in Appendix E. If there is demand, in future a method to directly measure the horizontal wavefield derivatives in the borehole will be created, and the method of calculating derivatives using Equation 6.12 emulates that situation. In the absence of such technology and of closely spaced boreholes, the approximation in Equation 6.13 can be used.

When using Marchenko-calculated wavefields an additional step is needed to scale the Green's functions to compensate for changes in amplitude with depth and offset after carrying out Marchenko but before applying the convolution of Equation 6.11. For a case of constant velocity but variable density each trace is normalized to the amplitude A at each depth and offset location. This depends on the downgoing transmission response τ^+ and the spherical divergence such that

$$A(x, z) = \frac{\tau^+(z)}{\sqrt{d(x, z)}} \tag{6.14}$$

where $d(x, z)$ is the distance between the surface source and virtual receiver. The downgoing transmission response is simply the cumulative transmission response of each layer through which the wavefield has passed. Since the transmission response at an interface is related to the zero-offset reflection response R_0 by $\tau = 1 - R_0$, the cumulative transmission response can be calculated from the reflection response. The latter can be determined down the vertical column of virtual receivers by the deconvolution of the Marchenko-calculated up and down going wavefields (Equations 6.3 and 6.4)

$$R_0(x, x, \omega) = \frac{G^-(x, x_1, \omega)}{G^+(x, x_1, \omega)} \tag{6.15}$$

Equation 6.14 applies only for a medium with constant velocity; for a case with variable velocity at short offsets from the vertical receivers the above equation would still hold approximately. Once the wavefields have been calculated and scaled if necessary, the inversion scheme of Equation 6.10 is implemented using

data from Equation 6.11 where

$$\begin{aligned}\mathbf{d} &\equiv \frac{\mathrm{d}}{\mathrm{d}t}\hat{R}(x_2, x_1, t)s(t)\\ \mathbf{A} &\equiv \left[\hat{G}(x, x_2, t)n_i\partial_i\hat{G}(x, x_1, t) - n_i\partial_i\hat{G}(x, x_2, t)\hat{G}(x, x_1, t)\right]\\ \mathbf{m} &\equiv \frac{1}{\rho(x)}\end{aligned} \tag{6.16}$$

It should be noted that the inversion scheme can also be implemented in the frequency domain. In this case matrix **A** would contain the multiplication of Green's functions for each frequency. If Marchenko wavefields are used the convolution result must be scaled to the reflectivity of the primary reflection prior to inversion. This is achieved by assuming the density of the top layer is known.

6.4 Results

6.4.1 Physical receivers in a borehole

The method was tested on a synclinal model using the density structure shown in Figure 6.3a and velocity structure shown in Figure 6.3b. A borehole was placed at a horizontal location of 1000m with two outside columns of receivers for the dipole calculation located at 998m and 1002m. Receivers were placed at 2m intervals down the borehole. Two arrays of 20 sources are placed on the surface either side of the borehole with 16m spacing between sources, the first at horizontal positions 640m-944m, the second at 1120m-1424m. To simulate recordings in a borehole, wavefields are modeled using an acoustic finite-difference code with absorbing boundary conditions on all sides.

The dipole responses were calculated using both the finite-difference method (Equation 6.12) and plane-wave approximation (Equation 6.13). The angle θ was calculated with the known velocity model in this case; for a more complex case a data driven method could be used to calculate the slope (e.g., Fomel (2007)). Figure 6.4 compares the geometries used by the two methods and the results. Both methods produce similar horizontal derivatives so either method can be used for this numerical example.

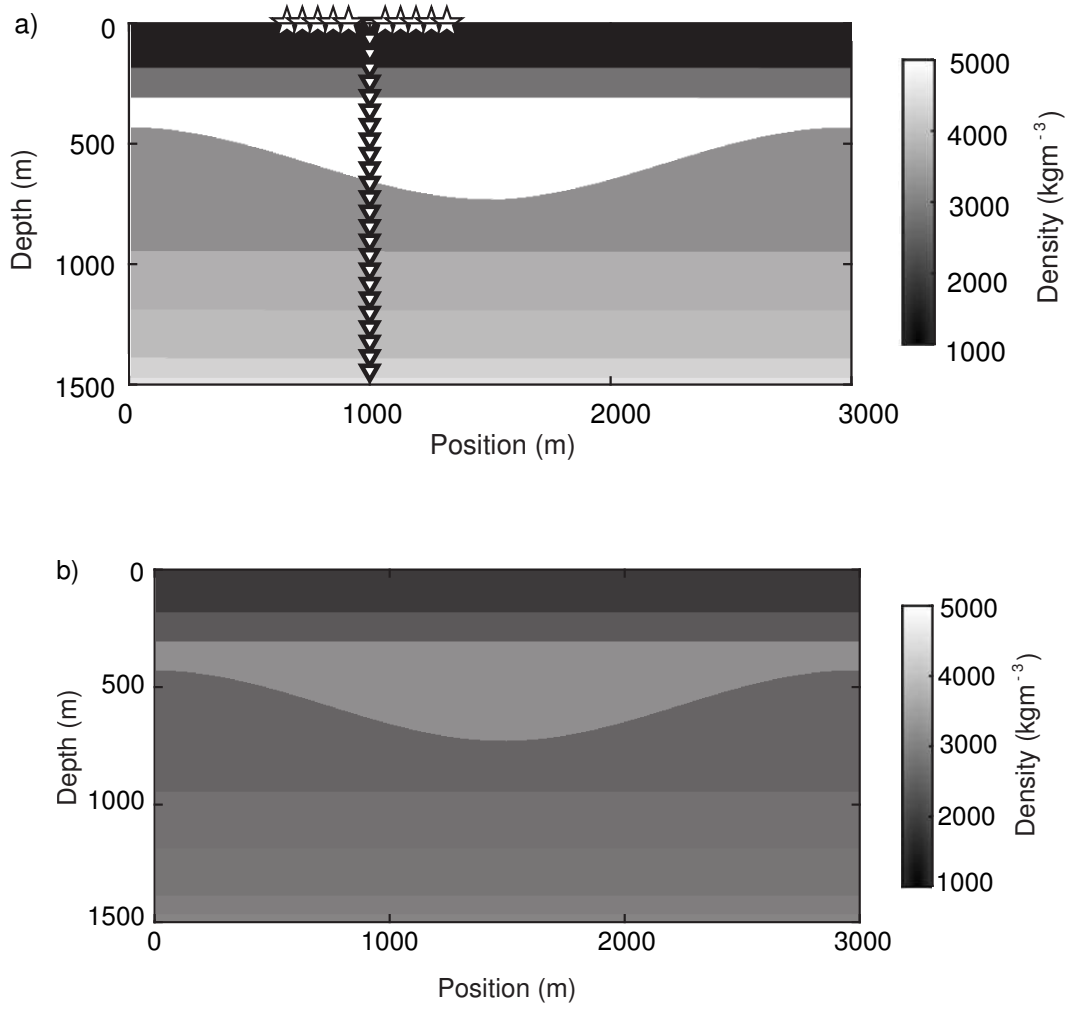

Figure 6.3 a) Density profile for the synclinal model and b) velocity profile for the synclinal model. Stars represent locations x_1 and x_2 of source arrays. Triangles represent subsurface downhole receivers along boundary $\partial\mathbb{D}$ in Figure 6.2.

All of the dipole responses were calculated using Equation 6.12, then the arrays of wavefields are convolved together according to Equation 6.11 to give 400 (x_1, x_2) pairs. Figure 6.5 compares the measured surface reflectivity with the convolution result for a surface source at 640m and surface sources acting as virtual receivers between 1280m-1424m with 16m spacing, and using the correct density along the borehole. The reflectivity is shown to be recovered correctly from Equation 6.11 when the density is known.

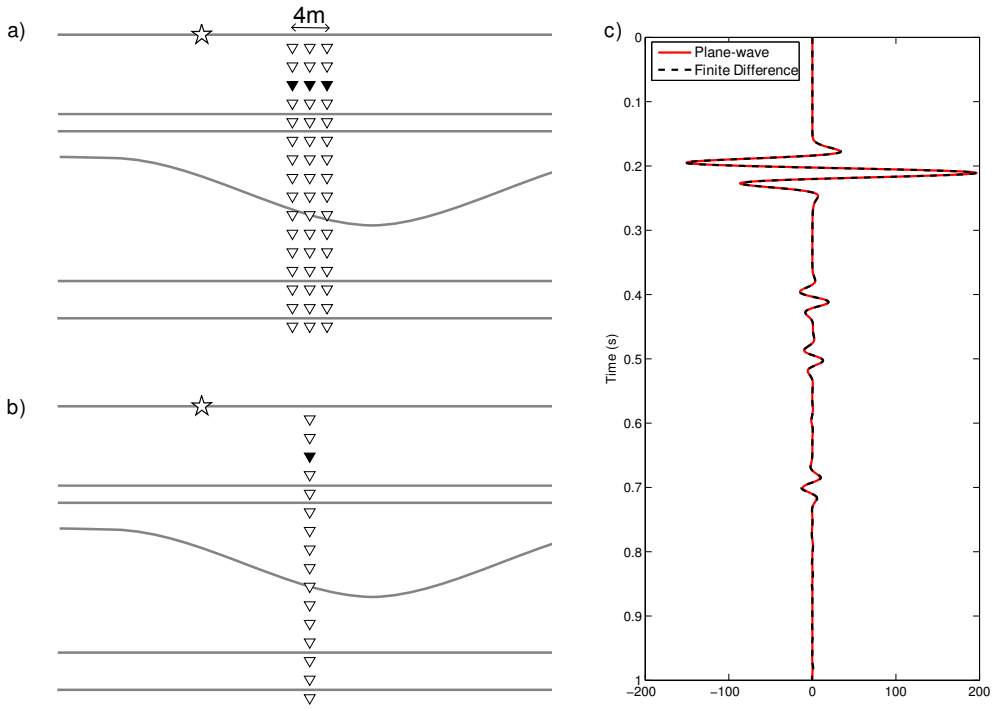

Figure 6.4 Schematic geometry for estimating the horizontal derivative or dipole response in a borehole. The star represents a surface source location, white triangles represent subsurface receiver locations and the black triangles are the locations of receivers used for the trace comparison in c). a) The two outer receivers (black) are used in Equation 6.12 to calculate the dipole response using the finite-difference method. b) The central receiver is used in Equation 6.13 to calculate the dipole response by plane-wave approximation. c) Trace comparison of dipole response for the finite-difference method (black dotted line) and plane-wave approximation (red line) from a source at 640m on the surface to a receiver at 1000m horizontally and a depth of 400m.

Then μ is calculated by the L-curve method and the full data set is inverted to give a value for $\frac{1}{\rho(x)}$ at each receiver point x in the borehole. Figure 6.6 shows the result of the inversion after converting values of $\frac{1}{\rho(x)}$ values to $\rho(x)$. Using borehole data to a depth of 1600m the subsurface density has been estimated well to a depth of around 1200m (below this depth the results depart significantly from the true density as shown in Figure 6.7).

The experiment was repeated for a series of laterally-shifted vertical boreholes placed 100m apart at horizontal locations 500m-1500m. Again using surface arrays of 20 sources either side of each borehole, the inversion was performed at each borehole location then interpolated across the section to create the resulting image

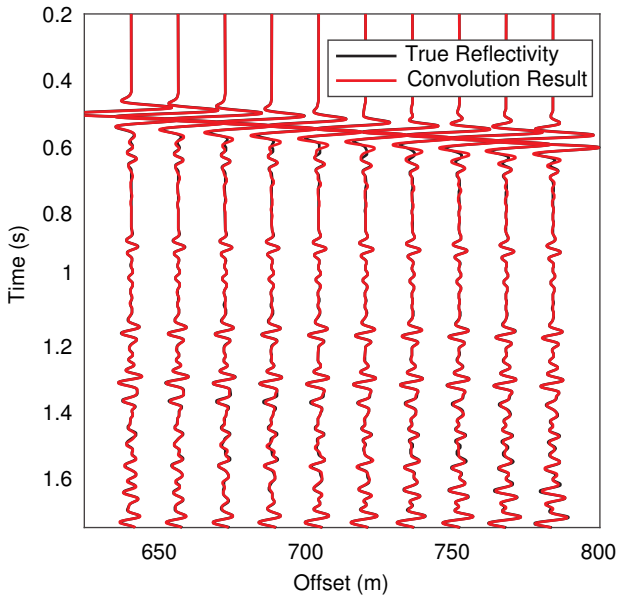

Figure 6.5 Trace comparison of measured reflectivity for a source at 640m to an array of virtual receivers at the surface between 1280m-1424m (black), with the result of the convolution (right side of Equation 6.11) calculated using the downhole receiver measurements in the geometry shown in Figure 6.3a (red).

in Figure 6.7. The scheme in Box 1 thus produces a model with a reasonably well defined density structure in the case where real receivers are used in boreholes.

6.4.2 Virtual receivers in the subsurface

The case is then tested where virtual receivers must be constructed in the subsurface using Marchenko methods. Using the same density and velocity model as in the previous example (Figure 6.3). Wavefields are calculated using the Marchenko method outlined above and described in detail in Wapenaar et al. (2014). The direct arrival estimate to be inserted into Equation 6.7 is calculated using an eikonal solver in the smoothed velocity model to obtain the travel time, and then the amplitude is set to unity; this method will be referred to as *standard Marchenko*. The Marchenko scheme used here requires that free surface multiples are not present, so the reflectivity is modeled for the medium using an acoustic finite-difference code with absorbing boundaries on all sides, and for co-located

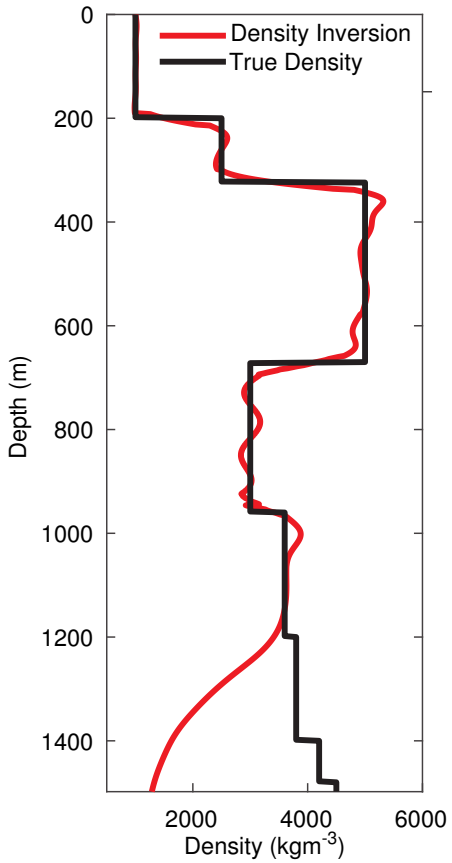

Figure 6.6 Inversion for vertical density profile at horizontal location 1000m. True density is shown in black, the density estimate using Equations 6.10 and 6.16 is in red.

sources and receivers placed at 8m intervals along the top surface of the model. Three vertical boundaries of virtual receivers are placed 2m apart with the central column at horizontal location 1000m (Figure 6.3a), and full Marchenko wavefields are calculated for each virtual receiver. Two surface arrays of 30 sources at 8m spacing are selected either side of each borehole. The first array is at positions 740-972m along the surface of the model, and the second array at positions 1020-1252m. The Marchenko derived wavefields from each surface array location to subsurface receivers are used to convolve the wavefields on either side of the vertical boundary using Equation 6.11 to give 900 (x_1, x_2) pairs. Marchenko wavefields do not have correct absolute amplitudes so they must be normalized to the reflectivity before the inversion can be performed. To achieve this I assume that the density of the top layer is known so that the convolution result can be

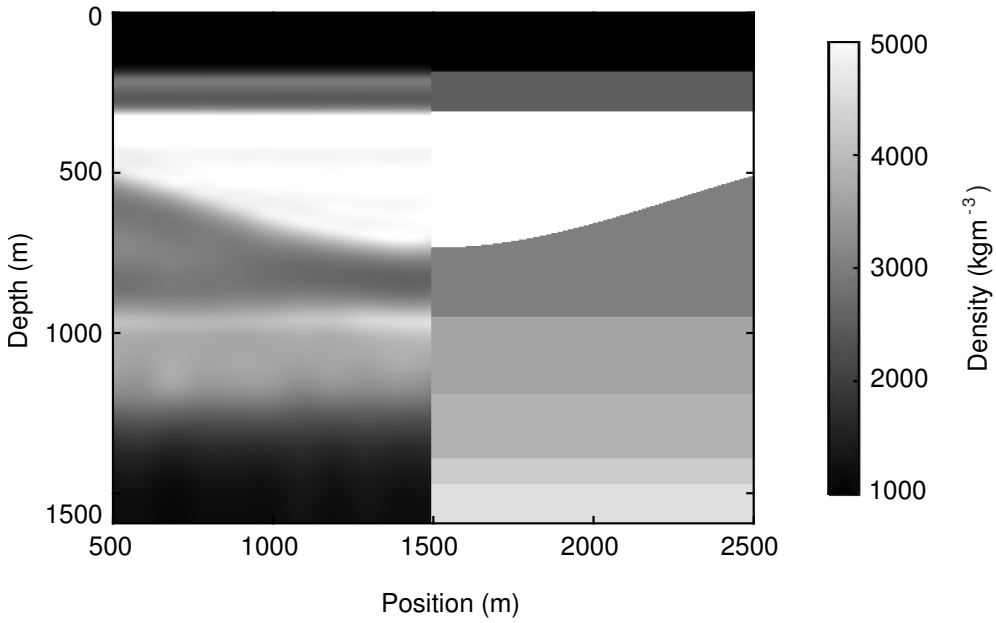

Figure 6.7 2D model for density. Left side of image is the inversion result, right side of image is the true model (which is laterally symmetric).

scaled to the reflectivity of the first (primary) reflection. The full data is then inverted for $\frac{1}{\rho(x)}$ using Equations 6.10 and 6.16.

Figure 6.8a shows the result of the convolution using Marchenko wavefields with the direct arrival calculated using a smooth velocity model, and this is compared with the measured reflectivity for a surface source at 900m and surface receivers between 1092-1064m with 8m spacing horizontally. All events are reconstructed in each trace, but the amplitudes of each event are not recovered correctly, as can clearly be seen between 0.5-0.8 seconds. This is due to the errors in amplitudes of the Green's function estimates for subsurface receivers that cause amplitude errors in the summation in Equation 6.11 (see the Discussion for more details). The inaccuracy in the amplitudes causes an erroneous inversion result for density as can be seen in Figure 6.9. Figure 6.9a shows the inversion result for $\frac{1}{\rho(x)}$ and Figure 6.9b the final density result $\rho(x)$ at 1000m along the model. Clearly, using this method alone we are not able to recover the correct density.

In order to obtain more accurate amplitudes from the Marchenko method it is necessary to calculate the variation of amplitude with depth for the Green's

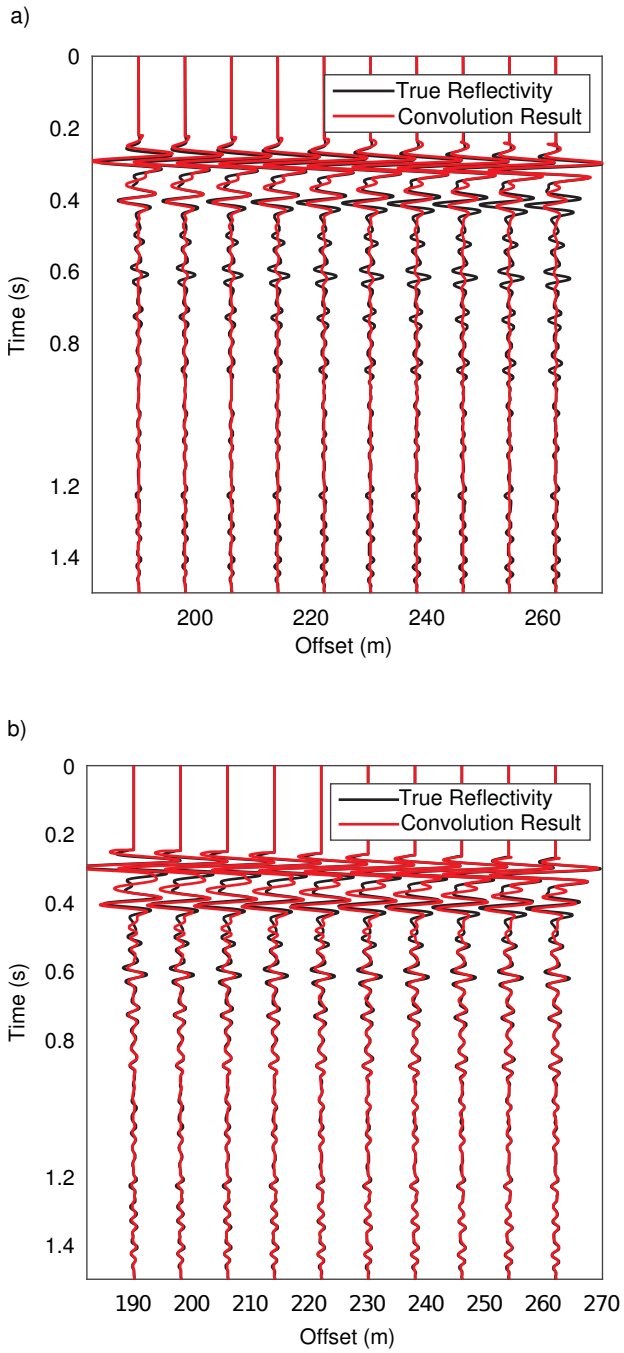

Figure 6.8 Trace comparison of measured reflectivity for a single source at location 900m horizontally to an array of surface receivers between 1092-1164m (black), with the result of the convolution in Equation 6.11 calculated using Green's functions estimated using the Marchenko equations (red). (a) shows traces from the Marchenko scheme when the direct arrival has an amplitude of 1. (b) shows traces from the Marchenko scheme when the direct arrival amplitude is set as if it had been measured in a borehole. Traces have been individually scaled as described in the text.

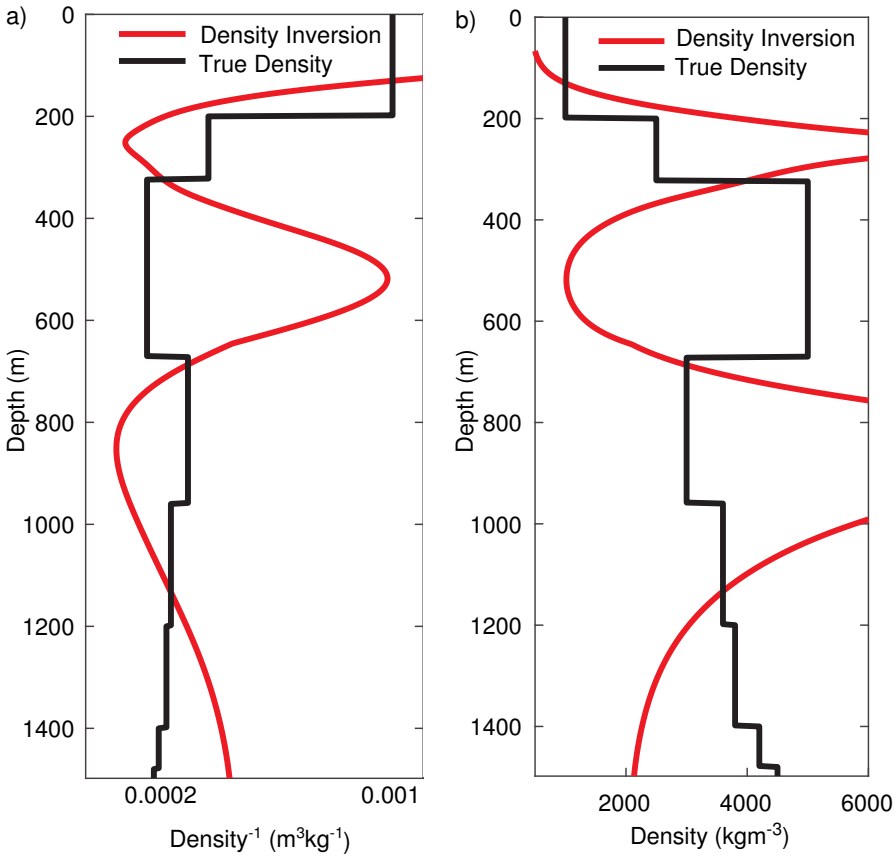

Figure 6.9 (a) Density inversion result for $\frac{1}{\rho(x)}$ when using an estimate of the direct arrival with amplitude 1 in the standard Marchenko method. (b) Reciprocal of curve in (a) giving the final density result $\rho(x)$, at horizontal location 1000m.

function estimates. Using Equations 6.14 and 6.15 the amplitude of the direct arrivals are calculated for each offset and depth location and then scaled each Marchenko calculated Green's function; this method will be referred to as *normalized Marchenko*. The same geometrical set up was used as above and the Marchenko-derived wavefields were used to apply Equations 6.10 and 6.16. Figure 6.8b shows the result of the convolution using Marchenko wavefields with amplitudes corrected using Equations 6.14 and 6.15. The amplitudes of events calculated by the convolution of normalized Marchenko traces are now more representative of the true amplitude of the events.

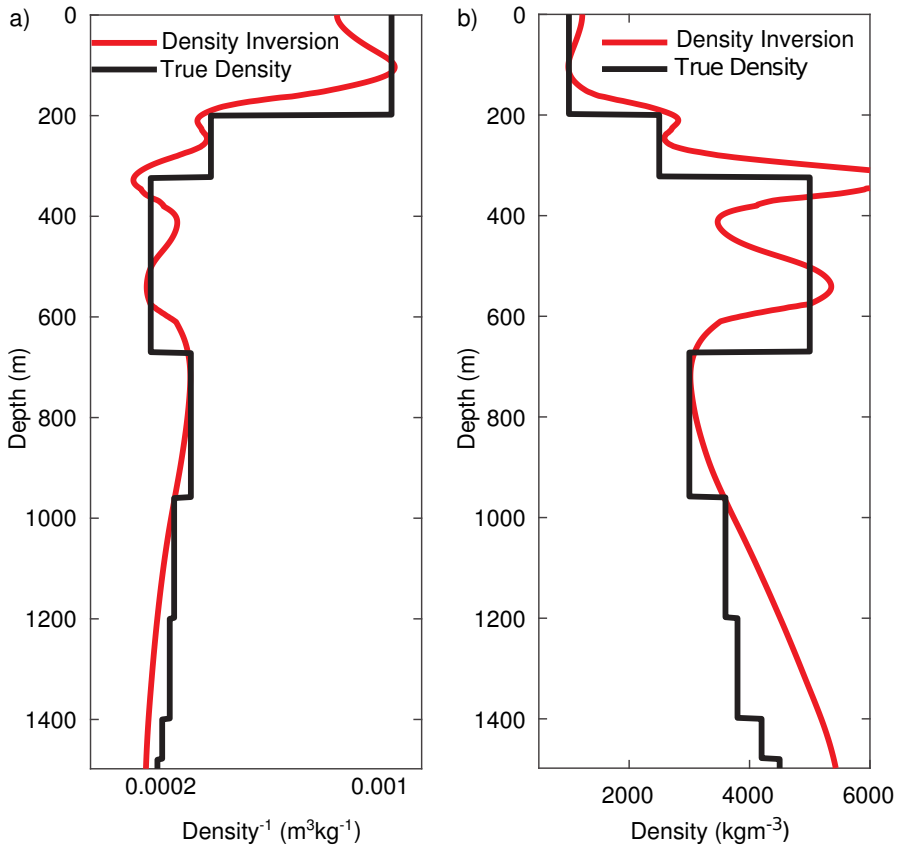

Figure 6.10 (a) Inversion result for $\frac{1}{\rho(x)}$ when using the normalised Marchenko method (see text for details). (b) Reciprocal of curve in (a) giving the final density result $\rho(x)$, at horizontal location 1000m.

Figure 6.10 shows the inversion result for $\frac{1}{\rho(x)}$ and the final density result $\rho(x)$ at 1000m along the model. Due to amplitude errors that remain in the Marchenko method and hence in the convolution result, the inversion is less accurate at depth than when using borehole data (Figure 6.6). Nevertheless, a reasonable density model can be obtained for the upper layers. This result shows that if the amplitude of the direct arrivals can be inferred for a vertical column of receivers, for example through estimation of the reflectivity coefficients, then this method to estimate density will work; a background density profile of the subsurface might then be obtained.

6.5 Discussion

Our method aims to exploit the linear relationship between wavefield data and (the reciprocal of) density in the convolutional interferometric Equations 6.2 and 6.11 using a vertical array of subsurface receivers, with the aim of being able to estimate density from only single-sided reflection data. The main alternative method to estimate density is examining the waveform amplitude variation with offset (AVO). AVO inversion requires direct inversion of long offset data, which tends to have a lower signal to noise ratio and be more sensitive to velocity errors than near-offset data, and it also requires the removal of free surface and internal multiples (Li, 2005) . Therefore the gathers must be preconditioned before inversion. By contrast, our method can be applied by inverting near offset data and no internal multiple removal is needed. it can therefore be implemented earlier in a work-flow (for example the density calculated here can be used as a background density for elastic FWI). In fact when using Marchenko calculated wavefields, the inversion of near offset data gives better results (discussed below).

When using wavefields measured directly in a borehole, this method is able to estimate the density in the subsurface down to around 1200m when using a borehole that extends to a depth of 1600m. Since this result is based on a truly linear inversion scheme, no prior model of density was needed. The method depends on the ability to calculate horizontal pressure gradients from a vertical borehole, given estimates of the wavefield azimuth at the receiver and the slowness, or to record them directly for which currently there is no practical standard technology. Moldoveanu et al. (2017) have recently shown data acquired using a three-dimensional array of hydrophones attached to an autonomous marine vehicle that enables the calculation of first and second spatial derivatives of the recorded pressure field in marine settings. However, as in our synthetic example above they require pressure arrays to be offset spatially in the direction of the derivative, which is usually not possible in borehole settings. In recent years portable instruments that are used at the surface to record rotational ground motions have been developed (Schmelzbach et al., 2018). Since rotational measurements can be related to the displacement gradient tensor wavefield, if such instruments were

made to be used inside a borehole setting this technology could be used to measure the horizontal pressure gradients in the future.

If instead of using wavefields recorded in a physical borehole we use wavefields from virtual receivers in the subsurface calculated by the standard Marchenko method, the results are poor. Amplitude errors in the final Green's functions render the inversion results unreliable. The amplitude errors in the final Green's functions are directly related to the estimate of the transmission response that is used to initiate the Marchenko iterative scheme. The direct arrival of the inverse of the transmission response (Equation 6.7) is usually approximated by the time reverse of the direct arrival of the Green's function, which is estimated using an eikonal equation solver and a smooth velocity model in standard Marchenko. In almost all published work to date, the estimated direct arrival is then scaled to have identical amplitude not only across each gather but also across each virtual receiver location with depth. In effect, all amplitude differences due to transmission responses and amplitude variations with offset are thus lost. For imaging applications using Marchenko, a deconvolution imaging condition is often applied (Wapenaar et al., 2014); the transmission response related errors will then be removed by division, which explains why the issue of absolute amplitude scaling has received relatively little attention up to now. Therefore the issues that arise in density estimation do so because no such division is included in Equation 6.11.

Vasconcelos et al. (2015) show how standard Marchenko has limitations in complex models, especially in areas with large parameter contrasts such as beneath salt. They formulate alternative initial focusing functions based on reference models such as conventional migration models, and normalize the initial focusing functions using impedance from the reference models. Using this method they achieve better estimates of subsurface wavefields, retrieving more internal multiple reflections and a better first arrival amplitude. However, their method relies on prior knowledge about exactly the amplitudes that are used to estimate density in our inversion scheme. More recently, van der Neut et al. (2018) solve a non-linear minimization problem to scale the amplitudes of a standard Marchenko direct arrival in 1D. Those authors comment that an extension to 2D and 3D would

likely involve an angle-dependent correction and should be the subject of further research. Suffice to say that scaling Marchenko results correctly in the real Earth remains an unsolved problem. The conclusion of this work is that in cases where we want to estimate wavefields that have the correct relative scaling with depth and offset, the true transmission responses need to be taken into account when calculating the direct arrival, or the scaling should be calculated and applied after the Marchenko scheme but before the convolution.

To understand why the recovery of amplitudes is so important for the convolution method let's take two wavefields $\hat{G}(x, x_1, t)$ and $\hat{G}(x, x_2, t)$ that are Marchenko-derived wavefields calculated using an estimate of the direct arrival with a smooth velocity model. If we define depth-dependent scaling factors $\alpha(x)$ and $\beta(x)$ for wavefields $\hat{G}(x, x_1, t)$ and $\hat{G}(x, x_2, t)$ respectively, that represent the amplitude error due to the transmission response at each depth down the vertical column of virtual receivers. Then Equation 6.11 can be written as

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\hat{R}(x_2, x_1, t)s(t)\\&= \int_{x\in\partial\mathbb{D}} \frac{1}{\rho(x)}\left[\beta(x)\hat{G}(x, x_2, t)\alpha(x)\partial_i\hat{G}(x, x_1, t) - \beta(x)\partial_i\hat{G}(x, x_2, t)\alpha(x)\hat{G}(x, x_1, t)\right] n_i\partial\mathbb{D}.\end{aligned} \tag{6.17}$$

Thus it can be seen that if there exist unknown amplitude errors $\alpha(x)$ and $\beta(x)$ and the convolutional method to estimate density is applied, we are effectively inverting for $\frac{\alpha(x)\beta(x)}{\rho(x)}$. The dependency of α and β with depth causes the summation over the vertical boundary to give amplitude differences between the Marchenko calculated trace and the true reflectivity, resulting in errors in density estimates. The problem can be simplified by assuming that there is no strong lateral heterogeneity in the model. Taking equal source offsets either side of the borehole it can then be assumed that $\alpha(x) = \beta(x)$ by Equation 6.14 and the problem is reduced to two unknowns $\alpha^2(x)$ and $\rho(x)$: by taking multiple traces with different offsets, a system of nonlinear equations can be set up and solved for $\alpha^2(x)$ and $\rho(x)$. There would still be one fewer equations than unknowns because $\alpha(x)$ would change with every new offset pair, so that the direct trade-off between $\alpha(x)$ and $\rho(x)$ would mean that the solution remains non-unique.

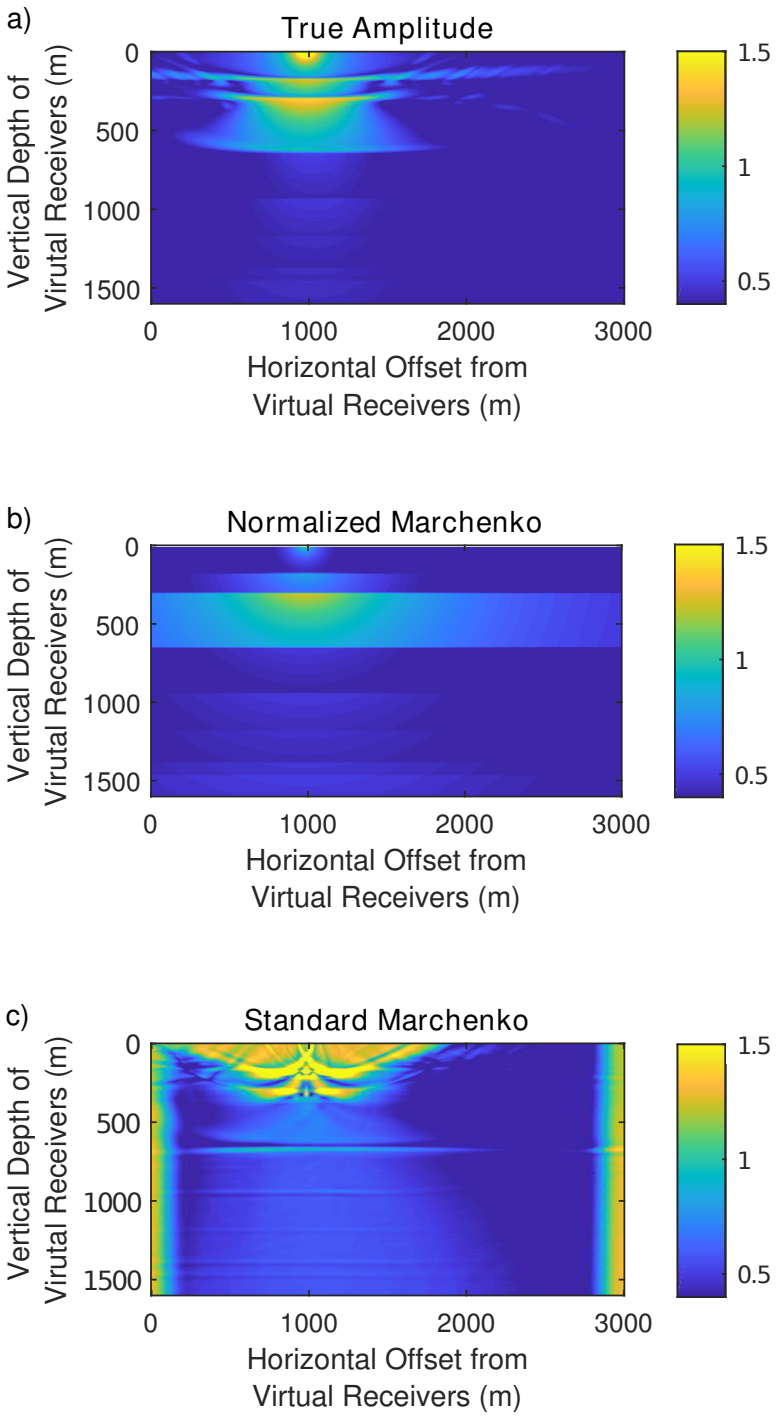

Figure 6.11 Images comparing normalized amplitudes of the first arrival of Marchenko-estimated Green's functions with that of the true Green's function between surface sources and the subsurface receivers shown in Figure 6.3a. a) Amplitude of true Green's function. b) Amplitude of Green's function estimated by normalized Marchenko. c) Amplitude of Green's function estimated by standard Marchenko. Each image shows maximum amplitude of a trace for sources with offset from the vertical column of virtual receivers on the horizontal axis, and receiver depth in the subsurface shown on the vertical axis.

In the numerical example above using virtual receivers, up- and down-going Marchenko Green's functions and the distance between the source and virtual receiver are used to calculate the variation of amplitude with offset using equations 6.14 and 6.15. Figure 6.11 compares the amplitudes of the Green's functions from standard and normalized Marchenko with the true Green's function for the model used in the numerical examples. When using standard Marchenko (Figure 6.11c) the resulting Green's function estimates do not represent the true Green's function amplitudes shown in Figure 6.11a. This explains the poor inversion result of Figure 6.9. However, the amplitudes of the normalized Marchenko Green's functions (Figure 6.11b) are much more realistic at near offsets but due to the constant velocity assumption of Equation 6.14 are still not accurate at far offsets. This is because the amplitude depends on the velocity as well as the transmission response and spherical divergence. Therefore, when selecting the near offsets for the convolution, the corresponding inversion result shown in Figure 6.10 is much improved but is still only able to give a smoothed density model. This can nevertheless be useful as an initial background model. Note that if the density and velocity of the top layer of the area of interest is known, a layer stripping method can be used to determine both the density and velocity of the medium. Using the background velocity model and the reflectivity calculated from Equation 6.15 the density and velocity can be calculated directly at each interface, as shown in Appendix F (Equation F.3). As with all layer stripping methods, the results will deteriorate with depth and this error would be compounded if the velocity or density varied within a layer.

In an exploration setting it may be possible to estimate the direct wavefield using a background velocity model and use a borehole to calibrate the amplitudes of the Marchenko Green's functions, then extrapolate these estimates laterally across the area of interest, for example using an image-guided approach (Hale, 2009). An alternative approach would be to invert the seismic data jointly with gravity data (which is directly sensitive to $\rho(x)$) to reduce uncertainties. However, as discussed in Blom et al. (2017), joint inversion of gravity and seismic data might not necessarily improve results without strong additional constraints due

to the inherent non-uniqueness and poor depth resolution in gravity inversions for density structure. This remains to be explored.

In summary, the method proposed has the potential to estimate density around a borehole, but is more complicated where no physical borehole exists. The complications arise directly from amplitude errors in the Marchenko method, which in turn originate from the fact that Marchenko methods depend on the amplitudes of the direct wave from the surface to every subsurface receiver. This chapter suggests a method to calculate the amplitudes, or infer density by layer stripping. However, improvements in this method are dependent on an area of current and future research: how to correctly normalize Marchenko estimates of Green's functions in the real Earth.

6.6 Conclusion

This chapter presented a new method to calculate density in the subsurface based on convolutional interferometry. The method was demonstrated on synthetic borehole acoustic data. We also show results when using virtual receivers in the subsurface, but practical limitations of the Marchenko method cause amplitude errors in Green's function estimation such that the amplitudes need to be calculated separately, or at least one borehole must still be present in the area of interest. The inversion is then also shown to be less accurate at depth compared with the case of using measured borehole data. Despite a constant velocity assumption to calculating the amplitudes the inversion is still possible but will only yield a background density model. However Marchenko methods represent an area of active research and future advances that better constrain the amplitudes of the estimated Green's functions are likely to directly improve the results of density inversion.

Chapter 7

Discussion

In this thesis I present novel techniques to recover subsurface density and velocity models from seismic data. The theory and techniques involved in probabilistic neural network inversion are introduced in Chapter 2, and Chapters 3, 4 and 5 demonstrate the application of this approach for 3 different scenarios. Chapter 3 uses neural networks to invert surface wave dispersion curves for shear-wave velocity profiles, I include uncertainty estimates of the dispersion curves in the network input to give more reliable mean velocity estimates and apply the networks to field data to produce shear-wave velocity maps at several depth levels. Chapter 4 demonstrates how neural networks can be used together with seismic gradiometry and wavefield inversion methods to produce depth-shear velocity maps for a near surface example. The full inversion process from field data to depth-velocity structure is computationally cheap, opening up the possibility of near-real time monitoring using dense arrays. Chapter 5 presents probabilistic neural network inversion of travel time data for 2D velocity maps and shows that informative and correct prior information is important in order to obtain mean and standard deviation estimates in high dimensional problems. Finally Chapter 6 moves away from neural networks to discuss a method to estimate subsurface density using seismic data without a direct trade-off with velocity, using a formulation of seismic interferometry that contains a linear dependency on density alone. I now discuss limitation and implications of this research for the fields of neural networks and Marchenko methods, and potentially fruitful future research directions.

7.1 Neural Networks

7.1.1 Dimensionality Reduction

In Chapter 5 I discussed the issue of training neural networks for high dimensional model parameter spaces. Since all sampling must be done prior to training, a large number of model-data pairs are needed in training sets to sample the high dimensional spaces adequately. For example, if we assume that any dimension in the model parameter space can be adequately described by only two samples, for an n-dimensional problem a training set would need 2^n samples to represent the space. The training would also be more computationally expensive as larger training sets might require larger network structures. Increasing the number of dimensions therefore rapidly renders the problem impractical. Chapter 5 showed that including more prior information in the models created for the training set improves the inversion results by restricting the training set and inversion results to a lower dimensional manifold and that including *correct* prior information is important for meaningful inversion results.

If there is a limited amount of geological or geophysical information about the subsurface then including this information in the training set might be insufficient to reduce a large dimensional problem to a lower dimensional manifold. In certain cases the number of dimensions in the problem could be in the order of 10^3 or 10^4 parameters so that alternative solutions must be found to reduce dimensionality. Autoencoder networks are neural networks that learn to copy the input to the output via an internal layer that describes a lower-dimensional representation of the input parameter space (Hinton and Salakhutdinov, 2006). The section of the network used to convert the input to the representation is called the 'encoder' and the section of the network used to convert the representation back to the original input is called the 'decoder' (Figure 7.1 Step 1). In the context of Geophysics, Valentine and Trampert (2012) discuss the use of auto-encoders for input parameter space dimensionality reduction. They reduce a waveform of 512-data points to a 32-element encoding that describes the original waveform, thus significantly reducing the dimensionality of the data. They suggest useful

applications could be to reduce the input space $\mathbf{d}$ thus requiring smaller network structures and therefore fewer network weights to optimise during training.

This does not solve our tomography problem directly as we would still need to estimate a large number of velocities to describe the Earth structure. However, it could be possible to train an auto-encoder network to reduce the dimensions of the input and output parameter spaces $\mathbf{d}$ and $\mathbf{m}$ to a lower dimension representation $\mathbf{L}_m$ and $\mathbf{L}_d$ respectively (Figure 7.1 Step 1). Then neural networks can be used to learn a mapping between the lower dimensional parameter spaces $\mathbf{L}_d$ and $\mathbf{L}_m$. Depending on the size of the input parameter space $\mathbf{d}$ this may not need to be reduced to a lower dimensional space $\mathbf{L}_d$ and can be used in its original form as an input to the network.

However, we would like not only an image but also the related uncertainties or, ideally, the full posterior probability density function. For this I propose an adjustment to the above scheme that can incorporate an MDN so the uncertainties of the final solution can be quantified, this is outlined in Figure 7.1. In Step 1 auto-encoders are trained to give representations $\mathbf{L}_d$ and $\mathbf{L}_m$ of the original input $\mathbf{d}$ and output data $\mathbf{m}$ respectively. An MDN is then used to learn the mapping between the lower dimensional spaces using the training set $T_L = \{\mathbf{L}_d, \mathbf{L}_m\}$ (Step 2) and then the outputs of the MDN are used to describe the joint distribution between the lower dimensional output parameters. Finally in Step 3, samples can be drawn from this joint distribution and used as an input to the decoder network to produce an ensemble of models that describe the possible velocity structures which can then be interrogated.

The quality of the velocity model produced from an auto-encoder network depends on how well the network is trained. Auto-encoders are not the only solution and the above method is just one example of how the problem could be solved. Future work in this area should try to understand the optimal way to parameterise velocity structures to reduce the problem from a grid structure to a lower dimensional mapping. This would also have wider applications for subsurface imaging problems such as Full Waveform Inversion (FWI).

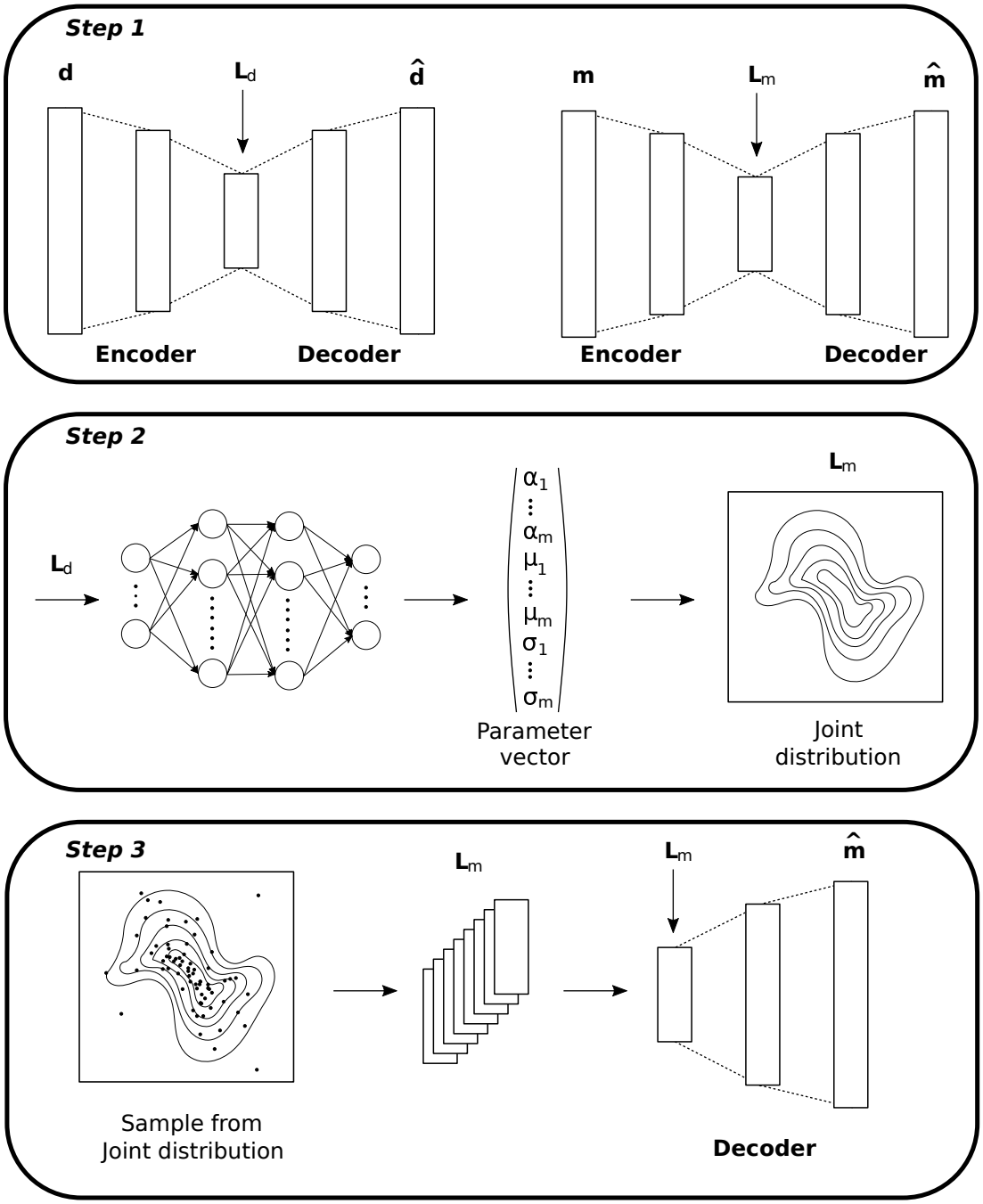

Figure 7.1 Illustration of proposed method for MDNs dimensionality reduction. In Step 1 auto-encoders are trained for the input and output parameter spaces $\mathbf{d}$ and $\mathbf{m}$ respectively, lower dimensional encoding representations $\mathbf{L}_d$ and $\mathbf{L}_m$ are created as well as a network that decodes these to the respective network input $\hat{\mathbf{d}}$ and $\hat{\mathbf{m}}$. In Step 2 an MDN is trained using the lower dimensional mappings $\mathbf{L}_d$ and $\mathbf{L}_m$ to provide the parameters α, μ and σ – the weights, means and standard deviations to provide a multidimensional Gaussian Mixture Model that describe the posterior space of $\mathbf{L}_m$. Step 3 samples from this mixture model to give a suite of examples for $\mathbf{L}_m$ which are then put into the decoder network to give an ensemble of examples that describe $\hat{\mathbf{m}}$.

7.1.2 Posterior Density Function Estimation

Methods that estimate full posterior pdfs such as Monte Carlo methods, are not often used in imaging problems due to the large computational cost. In Chapter 3 I show that for problems with many independent inversions, such as data from an oil field, neural networks can be more computationally efficient than Monte Carlo methods. The results in this thesis are generally comparable in quality to that of Monte Carlo methods for low-dimensional problems (for example Figure 3.15) however they are not a full replacement. Improvements to the density estimation that neural networks can provide for high dimensional spaces is a real opportunity to retrieve full posterior probabilities from non-linear inversion in near-real time. This is especially useful in industry applications with large data sets where methods are often linearized as a trade-off between speed and accuracy of result.

The easiest way to improve the method is to extend MDN's to invert for the full covariance matrix Σ as in Williams (1996) instead of the diagonal covariance matrix in Equation 2.9 so that the Gaussian density function estimated is

$$\Theta_i(\mathbf{m} \mid \mathbf{d}) = (2\pi)^{-c/2}|\Sigma(\mathbf{d})|^{-1/2} \exp\left\{-\frac{1}{2}(m_i - \mu_i(\mathbf{d}))^T \Sigma(\mathbf{d})^{-1}(m_i - \mu_i(\mathbf{d}))\right\} \tag{7.1}$$

where c is the dimensionality of $\mathbf{m}$. In this work Williams (1996) used one Gaussian kernel with the full covariance matrix in Equation 7.1 but this could be extended to more than one kernel. However, representing the full covariance matrix requires more parameters, increasing the number of network outputs from $(2c+1) \times M$ to $\frac{c(c+5)}{2} \times M$. For high dimensional problems this method becomes impractical as the number of network outputs increase.

However in computer science there is ongoing research into alternative methods for direct density estimation. The neural autoregressive distribution estimator (NADE) is an approach to model the distribution of high-dimensional vectors based on a type of network called a restricted Boltzmann machine (RBM) (Larochelle and Murray, 2011). This was extended by Uria et al. (2013) to give a joint density estimation of real-valued vectors. It decomposes the joint distribution

into a product of one-dimensional conditional distributions which are individually modelled using MDNs with shared internal parameters between each network. The autoregressive approach has been used in speech synthesis (van den Oord et al., 2016a) and to create tractable densities for image generation (van den Oord et al., 2016b) but has not been applied in the Geosciences. Using these machine learning algorithms could provide a viable way to produce joint posterior density functions for higher dimensional geophysical problems.

7.1.3 Neural Network Size

The neural network examples in this thesis involved problems with relatively small input and output dimensions along with forward problems that are relatively fast to compute. The largest data set used in Chapter 5 took less than 16 hours to create. The smaller sized networks in Chapters 3 and 4 were trained using CPUs on a compute cluster and the networks in Chapter 5 used GPUs in the same compute cluster. With the increase in compute power now readily available it would be possible to generalise the problems seen in Chapters 3 and 4 for data without fixed depth intervals, similar to the transdimensional inversions seen in Monte Carlo sampling (Bodin and Sambridge, 2009; Galetti et al., 2017). However, this would involve larger training sets as the network would need to invert for velocity and depth, thus increasing the prior model space. A training set in the order of 10-100 millions models would not be unreasonable, considering that the training set of 1 million models used in Chapter 3 required only 122Mb to store. However, care would need to be taken that this training set is representative of the prior. In this case my approach of generating random models within a specific distribution for each layer with relatively little correlation between velocities in each layer would not be optimal. A more sophisticated method of sampling from the prior model space would be needed to create a suitable training set efficiently.

A problem with increasing the training set size by increasing the size of the input and output parameters is that the size of the network (the number of nodes, weights and biases) would need to increase to be able to represent the relations between the two parameter spaces. This would increase the training time and

the memory needed both during the training and to store the trained networks post-training. The inversion of dispersion curves in this thesis used a relatively small amount of memory for training and model size and the models were trained on CPUs. For these examples it would be possible to increase the training set and model size and still give a reasonable training time by training these models on GPUs. However, the travel-time tomography example was trained on GPUs and used a larger amount of memory and more complex network structures. In this case, to increase data or model dimensionality, a method such as dimensionality reduction as discussed in Section 7.1.1 would need to be employed to keep these networks at a manageable size.

7.1.4 Neural Network Limitations

One of the most common problems levelled at machine learning techniques are that neural networks are a 'black-box': it is difficult to interpret the trained networks to understand their outputs. For many people this lack of ability to interpret the results leads to a mistrust of machine learning algorithms, especially neural networks. Just as a doctor would want to know why a machine learning algorithm has decided a patient has cancer, a geoscientist will want to know why the network has assigned that specific subsurface property. For neural networks to be used widely in industrial applications they either need to be interpretable so that users can understand the outputs of the network, or based more on the physics of the problem rather than a large amount of data so that users trust the network is giving them the correct answer. Neural networks with a probabilistic output, as presented in this work, are able to provide a low level interpretation of the results. A mean result may look correct, but if the uncertainty is large this shows the data provided for network training does not constrain the result; such an insight is unavailable from a standard neural network. However, this does not explain why the network produced that result and assigned that uncertainty in the first place.

Much research is on-going into explainable artificial intelligence (often called XAI). A method that has been designed for explanation is called Layer-wise

Relevance Propagation (LRP) (Montavon et al., 2017). This method identifies significant pixels in an image that contribute to a classification by performing a backward pass through the network. It has been used to show what features of an image networks are using to classify the image, and to highlight where networks are using unintended features (e.g., identifying images as horses because all horse images have the same watermark - (Lapuschkin et al., 2016)), as well as to understand speech signal classifications (Becker et al., 2018) and even for insights into quantum-chemical systems (Schütt et al., 2017). The question is, could this be extended to the regression-type problems seen in this thesis, to confirm whether the neural networks represent the true physics of a problem? For example, in the applications in Chapters 3 and 4, the networks invert dispersion curves for a velocity-depth structure. It is known that lower frequencies penetrate to greater depth in the Earth than higher frequencies (Aki and Richards, 2002) so we would expect that lower frequencies would have more importance for the results of the deeper depth layers than higher frequencies. If we could use LRP to confirm whether the trained networks have picked up this physics from the training set then we could have even more confidence in the results of the network.

Alternatively, one could embed physics into the neural network during training. This hybrid approach combines current physics driven models with machine learning and is called physics-guided neural networks. Two approaches are used: the first is to include physics in the training data set so that observed data values used to train the network are augmented with data derived from known physics-based models. All the networks in the thesis were trained on synthetic data derived from the physical equation we know govern Earth's subsurface properties, so it could be argued that these are physics-guided neural networks. However, a second approach is to use a physics-based loss function during training that punishes physically poor predictions. This will result in more reliable predictions as physically impossible results will be excluded.

7.1.5 Future of the field

With the ever increasing popularity of machine learning, evidenced by the large increase in abstract submissions at almost every Geoscience related conference, careful thought must be given to how best to leverage this interest in order to best use the technology to enhance or improve on current state of the art methods. The key to efficiently using machine learning techniques is to identify the applications where current methods are lacking such as simple tasks that could be automated, operations that are computationally expensive to perform, or areas where complex hidden patterns in the data could enhance our knowledge of the subsurface. It is clear from the applications in this thesis that one advantage of neural networks is for repeated inversions using the same trained network, rather than one time inversions. In our examples we propose that they can be used for monitoring of a field area where the networks were applied, however the networks were trained in a way that they were specific to a set of geological conditions that described the field area. By including physics and creating a training set of synthetic data the networks could be applied to different sets of geophysical data with similar geological properties. For example the networks from Chapter 3 could be applied to different North Sea reservoirs where the expected geology lies within the prior of the training data set. Thus operations that were computationally expensive such as nonlinear tomographic inversions could be applied very quickly.

To advance progress with machine learning in Geosciences one way would be to create benchmark data sets. These are standard data sets with which researchers can train and test their algorithms, evaluate the performance and check how well they perform compared to previous research. Often labelled data for supervised learning is hard to obtain or time-consuming to label, and advances in machine learning in general can partly be attributed to the availability of data sets on which new methods can be tested. For computer science there are multiple such data sets, the most famous being the MNIST handwritten digit data set (LeCun et al., 1998). There are many more data sets for a range of different tasks including facial or image recognition (Phillips et al., 1998; Deng et al., 2009), text data (Harper and Konstan, 2016) and speech data (Versteegh et al., 2015). One main

advantage of applying machine learning algorithms to a benchmark data set is model performance comparison: it is very difficult to compare the performance of different networks when each is applied to a different data set. By creating benchmark data sets Geoscientists would be able to easily compare which networks perform the best and adopt them in their own work. For industry related problems there are already synthetic data sets that people can use for benchmarking, such as the Marmousi model (Martin et al., 2002) for imaging problems for example, and Equinor's public data release of many different geophysical data from the Volve field (Equinor, 2018) is a major step forward. However, these are not data sets specifically made for machine learning algorithms and a large amount of time must be spent cleaning and organising the data before a project starts. Work on creating easily accessible, well-structured data sets in Earth Science would help advance progress in this field.

7.2 Marchenko

In Chapter 6 I outlined two methods for estimating density in the subsurface: a first method where direct recordings of the wavefield are available in the subsurface and a second method when no direct recordings are available and Marchenko redatuming provides these wavefields. With the second method the main issue encountered is that the amplitudes of the wavefields calculated with Marchenko are incorrect. In that chapter I suggest a method to calculate the amplitudes that works reasonably well, however Marchenko related research is an active field. Dukalski et al. (2019) proposed a scheme they termed 'augmented Marchenko' that combines energy conservation and minimum phase constraints to account for the lack of full bandwidth surface reflection data. Using the same method as augmented Marchenko Mildner et al. (2019) determine the angle dependent transmission losses to correct the direct arrival used to initiate the Marchenko scheme. The method proposed currently works only in laterally invariant media but as the authors state, retrieving Marchenko Green's functions with reliable amplitudes opens up the method to be used for target-orientated FWI with virtual

sources. It could also be used to provide Marchenko Green's functions with accurate amplitudes to improve the method to retrieve density in the subsurface.

It can be argued that a different method could be used to estimate subsurface density. Since the method proposed in Chapter 6 involves three closely spaced boundaries of receivers, the pressure data could be used to calculate the second order spatial derivatives and thus the gradiometry and wave equation inversion methods described in Section 4.2.1 could be used to estimate the local velocity. Using the reflectivity data the acoustic impedance can be calculated, then combining the estimated local velocity and the acoustic impedance the density can also be calculated at each receiver location. This method could be computationally cheaper as the convolution step in the inversion (Equation 6.16) is no longer needed. However, this method relies on the direct trade-off with an estimated velocity and impedance value. Any errors in these values will affect the final density result. In contrast the method presented in this thesis attempts to exploit a formulation of seismic interferometry that exhibits a dependency on density alone. Future research would benefit from a comparison of the two methods and determine the sensitivity of each method to velocity errors.

7.2.1 Application to surface waves

The application of density recovery in this thesis focussed on an exploration setting using body waves to estimate density in the subsurface. The main issue with the method when recorded wavefields from receivers within the subsurface are not available is the possibility of recovering these with accurate amplitudes. An alternative application would be to use surface waves with a dense array of receivers on the surface such as the acquisition geometry in Chapter 4. The advantage of this method would be that recordings are available at receivers on the surface. The geometry of the receivers is shown in Figure 7.2. The wavefields needed for the method can be attained through the cross-correlation of ambient noise recordings to give $N(x_1, x_2, t) = G(x_1, x_2, t) \otimes s_n(t)$ where $N(x_1, x_2, t)$ is the cross-correlated noise recording between receivers x_1 and x_2, $G(x_1, x_2, t)$ is the Green's function between x_1 and x_2, and $s_n(t)$ is the noise source term. The

convolutional interferometry expression for estimating density from a surface array, equivalent to Equation 6.11, becomes

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}N(x_2,x_1,t)\otimes s(t)\\&=\int_{x\in\partial\mathbb{D}}\frac{1}{\rho(x)}\left[N(x,x_2,t)\otimes\partial_i N(x,x_1,t)-\partial_i N(x,x_2,t)(x,x_1,t)\right]n_i\partial\mathbb{D},\end{aligned}\tag{7.2}$$

Assuming the source signature can be estimated, the left-hand side of Equation 7.2 can be calculated, and as long as the distance between receivers Δx is smaller than the minimum wavelength of the recordings so that derivatives $\partial_i N(t)$ can be calculated, then the same inversion scheme as outlined in Equations 6.16 can be used to estimate density. Any work on this method should test the reliability of the amplitudes recovered in $N(t)$, the estimation of the source signature, and the effect of spacing between receivers for the calculation of derivatives.

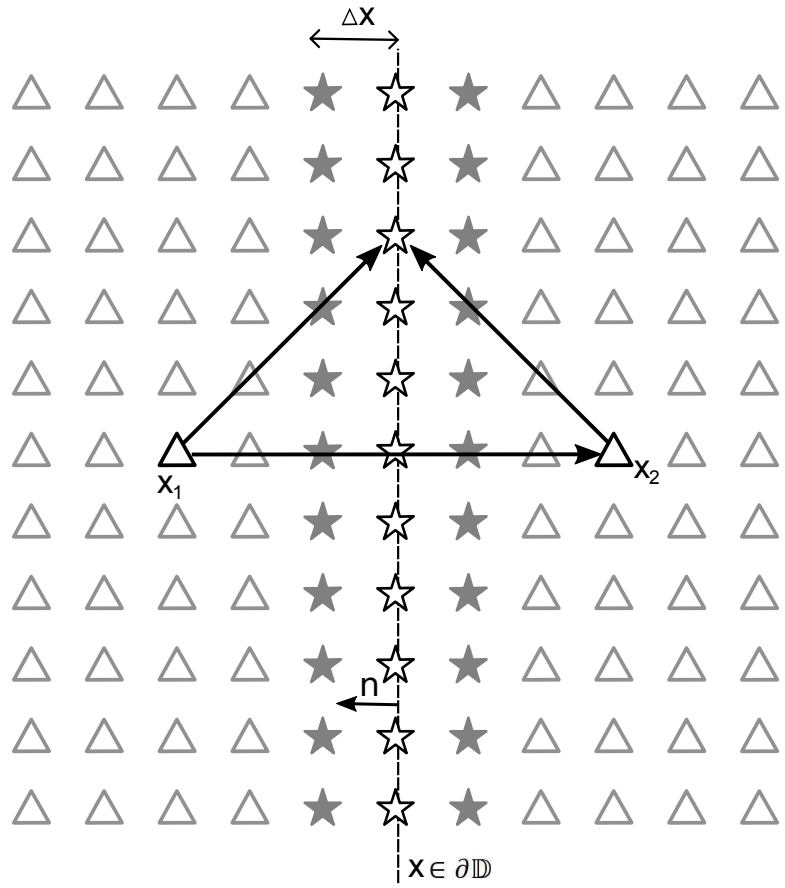

Figure 7.2 Schematic geometry for convolutional interferometry using surface waves to estimate density. The receiver array is laid out on the surface of the Earth. The boundary $\partial\mathbb{D}$ is represented by the black dashed line and virtual sources on the boundary are the central line of white stars. The neighbouring lines of grey stars are the virtual source locations used to calculate dipoles or derivative wavefields, and Δx is the horizontal distance between source lines. Receivers are represented by black triangles and black arrows represent example ray paths of acoustic wavefields. The grey stars represent the location of other available receivers in the array.

Chapter 8

Conclusion

Velocity and density estimates are essential in seismic imaging to produce accurate structural images of the subsurface. With the increase in compute power, machine learning techniques have become viable options for geophysical inversion techniques allowing rapid non-linear inversions using neural networks. The thesis explored this in detail using a modified standard form of neural network that produces a probability density function output so that uncertainties can be included in parameter estimates. This thesis also exploited the advent of Marchenko redatuming to produce wavefields that have been recorded in the subsurface to invert these wavefields for subsurface density estimates.

Mixture density networks were used to invert fundamental mode Rayleigh wave dispersion curves, calculated from Eikonal tomography, for subsurface shear-wave velocity structure beneath the Grane field in the North Sea. The neural network estimates were shown to be comparable to transdimensional Markov chain Monte Carlo results in 1D, 2D and 3D. It was shown that using including uncertainty estimates on the input data as a direct input to the mixture density networks provides more reliable estimates on noisy synthetic data.

This method was also demonstrated on near-surface field data. However the dispersion curves were estimated from wave equation inversion using a dense array of receivers instead of from Eikonal tomography . The wave equation inversion method did not provide uncertainty estimates on the dispersion curves, however the method was rapid and phase velocity dispersion curves were estimated

using short ambient noise recordings. The mixture density networks inverted the dispersion curves for subsurface shear wave velocity-depth models and these provided similar probability density functions to Monte Carlo sampling estimates. Wave equation inversion coupled with mixture density networks produced a full inversion process from field data to 3D depth-velocity structure which is fast, thus producing a technique for near-real time monitoring of the subsurface.

When dense arrays are not available wave equation inversion cannot be used to produce phase velocity estimates. Therefore if rapid velocity estimation is required travel-times should be inverted directly. This can be performed using mixture density networks for a fully non-linear travel time tomography inversion. Prior information in the training set is important for reliable velocity and uncertainty estimates: using more prior information can reduce the dimensionality of the problem so that the network can produce more accurate velocity results. However, if the prior information is false then the uncertainty estimates are incorrect. When the prior information is correct the mixture density networks exhibit uncertainty loops in the standard deviation maps, indicating that the networks appear to be able to emulate the correct physics of the tomography problem when the prior information was incorrect these loops disappear.

When inverting for density models a trade-off with velocity usually deteriorates the results. By exploiting the linear dependency on density in the convolutional interferometry equation it is possible to invert for density without a dependence on velocity. The method was demonstrated on synthetic data assuming receivers were available in the subsurface to measure the wavefields. Since this is not practical in field data situations the technique was also demonstrated using wavefields calculated using the Marchenko method with virtual receivers in the subsurface . The amplitude inaccuracies in the Marchenko wavefields rendered the inversion less stable than using measured wavefields, yet can still recover a background density model.

The results presented in this thesis show that using novel techniques such as machine learning can help to provide rapid estimation of material parameters used for subsurface imaging and estimates of the uncertainty of these parameters.

Also by incorporating more established techniques such as seismic interferometry with novel imaging methods such as Marchenko redatuming, new methods can be created to retrieve important material parameters that have been difficult to estimate previously. Both techniques are relatively in their infancy and I foresee that future research in all areas will extend the results presented here to become more influential in geophysics, not only for the uses in this thesis but for wider applications in geophysics.

Appendix A

Network configuration used in Tomography at Grane

The terminology used here is standard for neural networks and is defined succiently in Bishop (1995). The networks using Gaussian noise to simulate uncertainty in the data were trained using 3 fully connected layers (FC), where each node receives an input from every node in the previous layer. Between each node of the FC layers a rectified linear unit (ReLU) is used. The individual layer sizes and the total number of parameters to be trained in each network is outlined in Table A.1.

FC 1	FC 2	FC 3	Total parameters
200	300	200	133,145
400	200	350	173,545
400	1000	150	565,145
200	1000	350	570,745
400	500	350	398,845
400	1000	200	617,445
300	300	150	147,645
400	1000	350	774,345

Table A.1 Network configurations of the networks for which Gaussian noise of fixed standard deviation was added to the training set. Each network structure is trained 5 times with different random initialisations of starting parameter values.

Dispersion		Uncertainty	Total
FC 1	FC 2	FC 3	parameters
1295	240	500	573,045
1100	900	550	1,427,795
960	860	400	1,210,635
1000	220	1000	605,915
950	1000	140	1,300,315
1100	800	450	1,265,895
930	960	100	1,221,995
1200	200	600	517,295

Table A.2 Network configurations of the networks that included uncertainty vectors in the training set. Each network structure is trained 5 times with different random initialisations of starting parameter values.

The networks that included uncertainties as inputs were trained using 2 fully connected layers connected to the dispersion curve data and one fully connected layer connected to the uncertainty data, before concatenating the layers together and applying a further two hidden layers of size 250 and 150 respectively (Figure 3.4). In between each node of the fully connected layers a rectified linear unit (ReLU) is used. The individual layer sizes and the total number of parameters to be trained in each network is outlined in Table A.2.

Appendix B

Network configuration used in near-surface ambient noise inversion

The networks were trained using 3 fully connected layers (FC), where each node receives an input from every node in the previous layer. Between each node of the FC layers a rectified linear unit (ReLU) is used. The individual layer sizes and the total number of parameters to be trained in each network is outlined in Table B.1. For the final results the network with the lowest cost value of the test set is selected.

FC 1	FC 2	FC 3	Total parameters
100	50	100	11,559
10	100	100	12,159
100	100	100	21,609
100	100	50	16,109
50	100	100	16,359
100	10	50	2,519

Table B.1 Network configurations of the networks with 3 fully connected (FC) layers. Each network structure is trained 3 times with different random initialisations of starting parameter values.

Appendix C

Network configuration used in 2D Travel Time Tomography

The networks trained on individual cells used 4 fully connected layers (FC), where each node receives an input from every node in the previous layer. In between each node of the fully connected (FC) layers a rectified linear unit (ReLU) is used. The individual layer sizes and the total number of parameters to be trained in each network is outlined in Table C.1.

Size of model	FC 1	FC 2	FC 3	FC 4	Total No. of Parameters
8 x 8	270	1000	380	600	1,544,765
	100	500	450	550	1,622,685
	800	325	100	300	1,165,660
	200	400	200	50	334,335
	300	250	200	50	331,685
	900	700	70	550	2,077,505
	200	250	200	50	274,185
	300	400	200	50	406,835
16 x 16	375	500	300	600	5,265,470
	300	250	200	50	625,445
	200	400	200	50	628,095
	800	1000	500	550	6,076,995

Table C.1 Network configurations of the networks with 4 fully connected (FC) layers. Each row in the table represent a separate networks trained. Eight networks were trained for the 8 x 8 models and four networks for the 16 x 16 models.

Networks trained on the whole model (all cells at once) used a convolutional network with 3 convolutional layers (Conv) and 4 fully connected layers. The sizes of each layer and the total number of parameters to be trained in each networks is outlined in Table C.2.

Conv 1		Conv 2		Conv 3		FC 1	FC 2	FC 3	FC 4	Total No. of Parameters	
Filter	Kernel	Filter	Kernel	Filter	Kernel					8x8	16x16
128	5	128	5	64	1	800	150	600	1500	4,717,405	13,363,165
32	9	32	5	16	1	500	300	600	1500	4,354,183	12,999,943
32	9	32	5	16	1	500	200	2000	1250	5,641,438	12,847,243
32	9	8	5	16	1	500	300	600	1750	4,986,575	15,054,335
32	9	32	5	16	1	500	1500	50	1250	3,528,333	10,734,093

Table C.2 Network configurations of the convolutional networks with three convolutional (Conv) layers and 4 fully connected (FC) layers. Each row in the table represent a separate network trained.

Appendix D

Structural Similarity Index Measure (SSIM)

We use the form of the SSIM metric described in Wang et al. (2004). Let x and y be a window of NxN size. We calculate the luminance $l(x, y)$, contrast $c(x, y)$ and structure $s(x, y)$ defined as:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{D.1}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{D.2}$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{D.3}$$

where μ and σ are the mean and variance of the windows x or y and σ_{xy} is the covariance of x and y. To avoid instability in the division, constants C_1, C_2 and C_3 are defined as $C_1 = (k_1L)^2$ and $C_2 = (k_2L)^2$ where L is the dynamic range of the cell values while $k_1 = 0.01$ and $k_2 = 0.03$, and $C_3 = C_2/2$. The three components are combined to give the full SSIM:

$$SSIM(x, y) = [l(x, y)^\alpha \cdot c(x, y)^\beta \cdot s(x, y)^\gamma] \tag{D.4}$$

where α, β and γ are weighting parameters. Setting $\alpha = \beta = \gamma = 1$ we can simplify the expression to:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{D.5}$$

We perform the calculation over sliding windows and take the mean of the resulting $SSIM(x, y)$ values. For the 8 x 8 models we use 3x3 windows and the 16 x 16 models use 7x7 windows, so that the windows cover a similar spatial area.

Appendix E

Estimating horizontal pressure gradients in a borehole

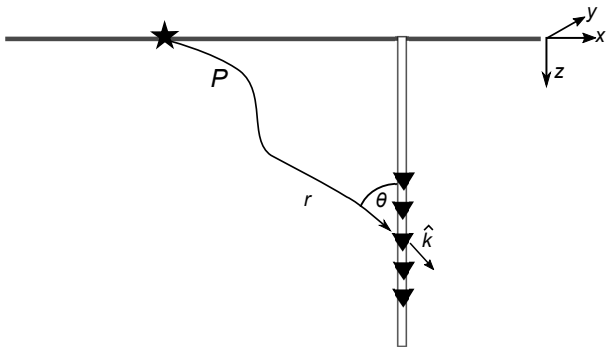

Figure E.1 Schematic diagram for calculating horizontal pressure gradient in a well. The source is represented by a black star, the receivers by black triangles, black arrow represents an example ray path of an acoustic wavefield, and r is the distance traveled in the $\hat{\mathbf{k}}$ direction, where k is the wavenumber.

To apply the convolutional interferometry equation of Equation 6.11 the horizontal pressure gradient $\partial_x G(x, x_1, \omega)$ across a borehole must be known. In practice this cannot be measured directly but can be estimated using the pressure measurement by the receiver in the borehole at depth z and time t using plane-wave approximation, a method to determine attributes of propagating seismic waves using the spatial derivatives of the wavefields. For our work we assume that there will be no out of plane reflections and therefore $\partial_y G(x, x_1, \omega) = 0$ at the borehole. We assume that between discontinuities a propagating wave

displacement, $P(t, x, z)$, takes the Cartesian form

$$P(t, x, z) = R(x, y) f(t - u_x(x - x_0) - u_z(z - z_0)), \tag{E.1}$$

where $f(t, x, z)$ is the phase variation of the wave in time and location, $R(x, y)$ is the geometrical spreading, u_x and u_z are the horizontal and vertical slownesses respectively and x_0 and z_0 and receiver locations, Figure E.1. Differentiating Equation E.1 gives (Langston, 2007c)

$$\frac{\partial P(t, x, z)}{\partial x} = A_x(x) P(t, x, y) + B_x(x) \frac{\partial P(t, x, y)}{\partial t}, \tag{E.2}$$

where

$$A_x(x) = \frac{\partial R(x, y)}{R(x, y)}, \tag{E.3}$$

$$B_x(x) = -\left[u_x(x) + \frac{\partial u_x}{\partial x}(x - x_0) \right]. \tag{E.4}$$

Assuming a plane wave propagating with azimuth θ (Figure E.1), the horizontal slowness $u_x(x)$ can be written as $u_r(r) \sin\theta$, where $u_r(r)$ is the slowness locally at the receiver location in the borehole. Since we are measuring the horizontal derivative at the borehole, $x = x_0$ and the second term in Equation E.4 can be ignored. Here $R(x, y) = 1$ and therefore $A_x(x) = 0$. Inserting these into Equation E.2 gives a relationship for the horizontal pressure gradient across a borehole dependent on the slowness, wavefield azimuth and time derivative of the pressure wavefield

$$\frac{\partial P(t, x, z)}{\partial x} = -u_r(r) \sin\theta \frac{\partial P(t, x, z)}{\partial t}. \tag{E.5}$$

Appendix F

Density and Velocity recovery for a variable velocity model

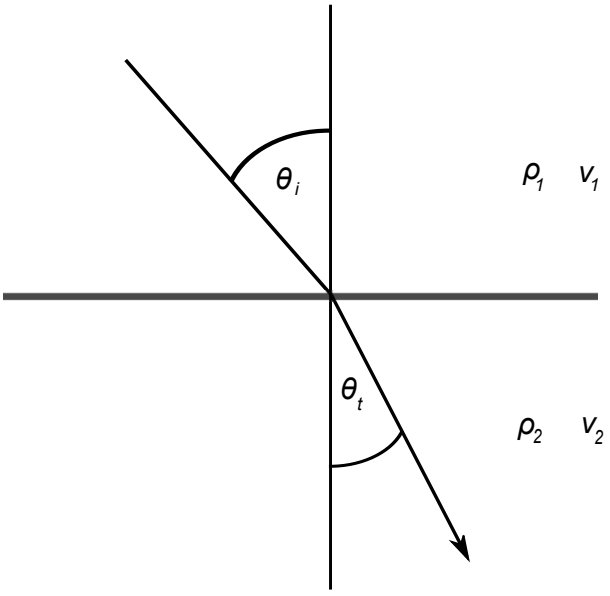

Figure F.1 Schematic diagram of the path of a wavefield at an interface. The angles θ_i and θ_t represent the incident and transmission angles, ρ is the density of the layer, v is the velocity and r is the reflectivity coefficient at the interface.

The Marchenko scheme uses a smooth velocity model, v_0 to estimate the direct arrival. At an interface, it can be assumed that

$$v_0 = \frac{v_1 + v_2}{2}, \tag{F.1}$$

where v_1 and v_2 are the velocities above and below the interface respectively (Figure F.1). The reflectivity coefficient r is defined as

$$r = \frac{\rho_2 v_2 - \rho_1 v_1}{\rho_2 v_2 + \rho_1 v_1}, \tag{F.2}$$

where ρ_1 and ρ_2 are the densities above and below the interface respectively. The reflectivity at each layer interface can be calculated from the up and down going Marchenko Green's functions, Equation 6.15, and if we assume the top layer velocity v_1 and density ρ_1 are known then, by a layer stripping method, we can use Equations F.1 and F.2 to determine the velocity and density of deeper layers with the following iterative scheme:

$$\begin{aligned} v_{i+1} &= 2v_{0,i} - v_i, \\ \rho_{i+1} &= \frac{-\rho_i v_i (1 + r_i)}{v_{i+1}(r_i - 1)}. \end{aligned} \tag{F.3}$$

Bibliography

Aki, K., 1957, Space and time spectra of stationary stochastic waves, with special reference to microtremors: Bull. Earthq. Res. Inst., **35**, 415–456.

Aki, K., A. Christoffersson, and E. S. Husebye, 1977, Determination of the three-dimensional seismic structure of the lithosphere: Journal of Geophysical Research, **82**, 277–296.

Aki, K., and W. Lee, 1976, Determination of three-dimensional velocity anomalies under a seismic array using first P arrival times from local earthquakes: 1. A homogeneous initial model: Journal of Geophysical research, **81**, 4381–4399.

Aki, K., and P. G. Richards, 2002, Quantitative seismology.

Allmark, C., A. Curtis, E. Galetti, and S. de Ridder, 2018, Seismic attenuation from ambient noise across the north sea Ekofisk permanent array: Journal of Geophysical Research: Solid Earth, **123**, 8691–8710.

Alwon, S., 2018, Generative adversarial networks in seismic data processing, *in* SEG Technical Program Expanded Abstracts 2018: Society of Exploration Geophysicists, 1991–1995.

Araya-Polo, M., T. Dahlke, C. Frogner, C. Zhang, T. Poggio, and D. Hohl, 2017, Automated fault detection without seismic processing: The Leading Edge, **36**, 208–214.

Araya-Polo, M., J. Jennings, A. Adler, and T. Dahlke, 2018, Deep-learning tomography: The Leading Edge, **37**, 58–66.

Aristodemou, E., C. Pain, C. De Oliveira, T. Goddard, and C. Harris, 2005, Inversion of nuclear well-logging data using neural networks: Geophysical Prospecting, **53**, 103–120.

Bai, J., and D. Yingst, 2014, Simultaneous inversion of velocity and density in time-domain full waveform inversion, *in* SEG Technical Program Expanded Abstracts 2014: Society of Exploration Geophysicists, 922–927.

Bakulin, A., and R. Calvert, 2006, The virtual source method: Theory and case study: Geophysics, **71**, SI139–SI150.

Becker, S., M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, 2018, Interpreting and explaining deep neural networks for classification of audio signals: arXiv preprint arXiv:1807.03418.

Bensen, G., M. Ritzwoller, and N. M. Shapiro, 2008, Broadband ambient noise surface wave tomography across the united states: Journal of Geophysical Research: Solid Earth, **113**.

Bensen, G., M. Ritzwoller, and Y. Yang, 2009, A 3-d shear velocity model of the crust and uppermost mantle beneath the united states from ambient seismic noise: Geophysical Journal International, **177**, 1177–1196.

Berbellini, A., A. Morelli, and A. M. Ferreira, 2016, Ellipticity of Rayleigh waves in basin and hard-rock sites in Northern Italy: Geophysical Journal International, **206**, 395–407.

Bergstra, J., B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox, 2015, Hyperopt: a python library for model selection and hyperparameter optimization: Computational Science & Discovery, **8**, 014008.

Bianco, M. J., and P. Gerstoft, 2018, Travel time tomography with adaptive dictionaries: IEEE Transactions on Computational Imaging, **4**, 499–511.

Bishop, C. M., 1994, Mixture density networks.

———, 1995, Neural networks for pattern recognition: Oxford University Press.

Bishop, T., K. Bube, R. Cutler, R. Langan, P. Love, J. Resnick, R. Shuey, D. Spindler, and H. Wyld, 1985, Tomographic determination of velocity and depth in laterally varying media: Geophysics, **50**, 903–923.

Blakely, R. J., 1995, Potential theory in gravity and magnetic applications: Cambridge University Press.

Blom, N., C. Boehm, and A. Fichtner, 2017, Synthetic inversions for density using seismic and gravity data: Geophysical Journal International, **209**, 1204–1220.

Bodin, T., and M. Sambridge, 2009, Seismic tomography with the reversible jump algorithm: Geophysical Journal International, **178**, 1411–1436.

Bodin, T., M. Sambridge, H. Tkalčić, P. Arroucau, K. Gallagher, and N. Rawlinson, 2012, Transdimensional inversion of receiver functions and surface wave dispersion: Journal of Geophysical Research: Solid Earth, **117**.

Borah, K., S. Rai, K. Prakasam, S. Gupta, K. Priestley, and V. Gaur, 2014, Seismic imaging of crust beneath the Dharwar Craton, India, from ambient noise and teleseismic receiver function modelling: Geophysical Journal International, **197**, 748–767.

Bregman, N., R. Bailey, and C. Chapman, 1989, Crosshole seismic tomography: Geophysics, **54**, 200–215.

Brocher, T. M., 2005, Empirical relations between elastic wavespeeds and density in the earth's crust: Bulletin of the seismological Society of America, **95**, 2081–2092.

Broggini, F., R. Snieder, and K. Wapenaar, 2012, Focusing the wavefield inside an unknown 1D medium: Beyond seismic interferometry: Geophysics, **77**, A25–A28.

Bussat, S., and S. Kugler, 2011, Offshore ambient-noise surface-wave tomography above 0.1 Hz and its applications: The Leading Edge, **30**, 514–524.

Calderón-Macías, C., M. K. Sen, and P. L. Stoffa, 2000, Artificial neural networks for parameter estimation in geophysics: Geophysical Prospecting, **48**, 21–47.

Campillo, M., and A. Paul, 2003, Long-range correlations in the diffuse seismic coda: Science, **299**, 547–549.

Cao, R., S. Earp, S. A. de Ridder, A. Curtis, and E. Galetti, 2019 In Press, Near-surface 3D seismic velocity models by wavefield gradiometry and neural network inversion: Geophysics.

Capes, T., P. Coles, A. Conkie, L. Golipour, A. Hadjitarkhani, Q. Hu, N. Huddleston, M. Hunt, J. Li, M. Neeracher, K. Prahallad, T. Raitio, R. Rasipuram, G. Townsend, B. Williamson, D. Winarsky, Z. Wu, and H. Zhang, 2017, Siri on-device deep learning-guided unit selection text-to-speech system: Proc. Interspeech 2017, 4011–4015.

Carney, M., P. Cunningham, J. Dowling, and C. Lee, 2005, Predicting probability distributions for surf height using an ensemble of mixture density networks: Proceedings of the 22nd international conference on Machine learning, ACM, 113–120.

Cary, P., and C. Chapman, 1988, Automatic 1-d waveform inversion of marine seismic refraction data: Geophysical Journal International, **93**, 527–546.

Castagna, J. P., M. L. Batzle, and R. L. Eastwood, 1985, Relationships between compressional-wave and shear-wave velocities in clastic silicate rocks: Geophysics, **50**, 571–581.

Chaves, C. A. M., and N. Ussami, 2013, Modeling 3-d density distribution in the mantle from inversion of geoid anomalies: Application to the yellowstone province: Journal of Geophysical Research: Solid Earth, **118**, 6328–6351.

Cho, K.-H., R. B. Herrmann, C. Ammon, and K. Lee, 2007, Imaging the upper crust of the korean peninsula by surface-wave tomography: Bulletin of the seismological Society of America, **97**, 198–207.

Choi, S., K. Lee, S. Lim, and S. Oh, 2018, Uncertainty-aware learning from demonstration using mixture density networks with sampling-free variance modeling: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 6915–6922.

Claerbout, J., 1985, Imaging the Earth's Interior: Blackwell Scientific Publications, Inc.

Claerbout, J. F., 1968, Synthesis of a layered medium from its acoustic transmission response: Geophysics, **33**, 264–269.

Connolly, P., 1999, Elastic impedance: The Leading Edge, **18**, 438–452.

Curtis, A., R. Cao, S. Earp, X. Zhang, S. de Ridder, and E. Galetti, 2019, Near-real time 3D seismic velocity and uncertainty models from ambient noise, gradiometry and neural network inversion: Presented at the 81st EAGE Conference and Exhibition 2019 Workshop Programme.

Curtis, A., P. Gerstoft, H. Sato, R. Snieder, and K. Wapenaar, 2006, Seismic interferometry-turning noise into signal: The Leading Edge, **25**, 1082–1092.

Curtis, A., and D. Halliday, 2010, Source-receiver wave field interferometry: Physical Review E, **81**, 046601.

Curtis, A., and A. Lomax, 2001, Prior information, sampling distributions, and the curse of dimensionality: Geophysics, **66**, 372–378.

Curtis, A., and J. O. Robertsson, 2002, Volumetric wavefield recording and wave equation inversion for near-surface material properties: Geophysics, **67**, 1602–1611.

Curtis, A., J. Trampert, R. Snieder, and B. Dost, 1998, Eurasian fundamental mode surface wave phase velocities and their relationship with tectonic structures: Journal of Geophysical Research: Solid Earth, **103**, 26919–26947.

Curtis, A., and R. Wood, 2004, Geological prior information: informing science and engineering: Geological Society of London.

Curtis, A., and J. H. Woodhouse, 1997, Crust and upper mantle shear velocity structure beneath the tibetan plateau and surrounding regions from interevent surface wave phase velocity inversion: Journal of Geophysical Research: Solid Earth, **102**, 11789–11813.

da Costa Filho, C. A., G. A. Meles, and A. Curtis, 2017, Elastic internal multiple analysis and attenuation using Marchenko and interferometric methods: Geophysics, **82**, Q1–Q12.

da Costa Filho, C. A., M. Ravasi, and A. Curtis, 2015, Elastic P-and S-wave autofocus imaging with primaries and internal multiples: Geophysics, **80**, S187–S202.

da Costa Filho, C. A., M. Ravasi, A. Curtis, and G. A. Meles, 2014, Elastodynamic Green's function retrieval through single-sided Marchenko inverse scattering: Physical Review E, **90**, 063201.

Dahlen, F., and J. Tromp, 1998, Theoretical global seismology: Princeton university press.

de Ridder, S., and B. Biondi, 2015, Near-surface scholte wave velocities at Ekofisk from short noise recordings by seismic noise gradiometry: Geophysical Research Letters, **42**, 7031–7038.

de Ridder, S., and A. Curtis, 2017, Seismic gradiometry using ambient seismic noise in an anisotropic Earth: Geophysical Journal International, **209**, 1168–1179.

de Ridder, S., and J. Dellinger, 2011, Ambient seismic noise eikonal tomography for near-surface imaging at Valhall: The Leading Edge, **30**, 506–512.

De Wit, R. W., A. P. Valentine, and J. Trampert, 2013, Bayesian inference of Earth's radial seismic structure from body-wave traveltimes using neural networks: Geophysical Journal International, **195**, 408–422.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, 2009, Imagenet: A large-scale hierarchical image database: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 248–255.

Devilee, R. J. R., A. Curtis, and K. Roy-Chowdhury, 1999, An efficient, probabilistic neural network approach to solving inverse problems: Inverting surface wave velocities for eurasian crustal thickness: Journal of Geophysical Research: Solid Earth, **104**, 28841–28857.

Dietterich, T. G., 2000, Ensemble methods in machine learning: International workshop on multiple classifier systems, Springer, 1–15.

Dębski, W., and A. Tarantola, 1995, Information on elastic parameters obtained from the amplitudes of reflected waves: Geophysics, **60**, 1426–1436.

Dosovitskiy, A., and T. Brox, 2016, Generating images with perceptual similarity metrics based on deep networks: Advances in neural information processing systems, 658–666.

Downton, J. E., and L. R. Lines, 2004, Three term AVO waveform inversion: SEG Technical Program Expanded Abstracts 2004, Society of Exploration Geophysicists, 215–218.

Duguid, C., D. Halliday, and A. Curtis, 2011, Source-receiver interferometry for seismic wavefield construction and ground-roll removal: The Leading Edge, **30**, 838–843.

Dukalski, M., E. Mariani, and K. de Vos, 2019, Handling short-period scattering using augmented Marchenko autofocusing: Geophysical Journal International, **216**, 2129–2133.

Dupont, E., T. Zhang, P. Tilke, L. Liang, and W. Bailey, 2018, Generating realistic geology conditioned on physical measurements with generative adversarial networks: arXiv preprint arXiv:1802.03065.

Dziewonski, A., A. Hales, and E. Lapwood, 1975, Parametrically simple earth models consistent with geophysical data: Physics of the Earth and Planetary Interiors, **10**, 12–48.

Dziewonski, A. M., and D. L. Anderson, 1981, Preliminary reference earth model: Physics of the earth and planetary interiors, **25**, 297–356.

Dziewonski, A. M., B. H. Hager, and R. J. O'Connell, 1977, Large-scale heterogeneities in the lower mantle: Journal of Geophysical Research, **82**, 239–255.

Dziewonski, A. M., and J. H. Woodhouse, 1987, Global images of the Earth's interior: Science, **236**, 37–48.

Earp, S., and A. Curtis, 2019a, Neural network travel-time tomography: Presented at the 81st EAGE Conference and Exhibition 2019 Workshop Programme.

———, 2019b, Probabilistic neural-network based 2D travel time tomography: arXiv preprint arXiv:1907.00541.

Earp, S., A. Curtis, S. Singh, and G. Meles, In Preparation, Estimating subsurface density by full waveform inversion of acoustic reflections using interferometric and Marchenko methods.

Earp, S., A. Curtis, X. Zhang, and F. Hansteen, Submitted, Probabilistic neural network tomography across Grane Field (North Sea) from surface wave dispersion data: Geophysical Journal International.

Edme, P., and S. Yuan, 2016, Local dispersion curve estimation from seismic ambient noise using spatial gradients: Interpretation, **4**, SJ17–SJ27.

Equinor, 2018, Disclosing all Volve data: https://www.equinor.com/en/news/14jun2018-disclosing-volve-data.html. (Accessed: 31-08-2019).

Farra, V., and R. Madariaga, 1988, Non-linear reflection tomography: Geophysical Journal International, **95**, 135–147.

Fishwick, S., B. Kennett, and A. Reading, 2005, Contrasts in lithospheric structure within the Australian craton—insights from surface wave tomography: Earth and Planetary Science Letters, **231**, 163–176.

Fokkema, J., and P. M. van den Berg, 1993, Seismic applications of acoustic reciprocity: Elsevier Science Publishing Company, Inc.

Fomel, S., 2007, Velocity-independent time-domain seismic imaging using local event slopes: Geophysics, **72**, S139–S147.

Friederich, W., 2003, The S-velocity structure of the East Asian mantle from inversion of shear and surface waveforms: Geophysical Journal International, **153**, 88–102.

Galetti, E., A. Curtis, B. Baptie, D. Jenkins, and H. Nicolson, 2017, Transdimensional love-wave tomography of the British Isles and shear-velocity structure of the East Irish Sea Basin from ambient-noise interferometry: Geophysical Journal International, **208**, 36–58.

Galetti, E., A. Curtis, G. A. Meles, and B. Baptie, 2015, Uncertainty loops in travel-time tomography from nonlinear wave physics: Physical review letters, **114**, 148501.

Gardner, G., L. Gardner, and A. Gregory, 1974, Formation velocity and density - The diagnostic basics for stratigraphic traps: Geophysics, **39**, 770–780.

Gerstoft, P., K. G. Sabra, P. Roux, W. Kuperman, and M. C. Fehler, 2006, Green's functions extraction and surface-wave tomography from microseisms in southern california: Geophysics, **71**, SI23–SI31.

Glorot, X., and Y. Bengio, 2010, Understanding the difficulty of training deep feedforward neural networks: Proceedings of the thirteenth international conference on artificial intelligence and statistics, 249–256.

Goodfellow, I., Y. Bengio, and A. Courville, 2016, Deep learning: MIT press.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, 2014, Generative adversarial nets: Advances in neural information processing systems, 2672–2680.

Graves, A., 2013, Generating sequences with recurrent neural networks: arXiv preprint arXiv:1308.0850.

Green, P. J., 1995, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination: Biometrika, **82**, 711–732.

Green, P. J., and D. I. Hastie, 2009, Reversible jump MCMC: Genetics, **155**, 1391–1403.

Guo, Z., Y. J. Chen, J. Ning, Y. Feng, S. P. Grand, F. Niu, H. Kawakatsu, S. Tanaka, M. Obayashi, and J. Ni, 2015, High resolution 3-D crustal structure beneath NE China from joint inversion of ambient noise and receiver functions using NECESSArray data: Earth and Planetary Science Letters, **416**, 1–11.

Gupta, S., K. Kothari, M. V. de Hoop, and I. Dokmanić, 2018, Deep mesh projectors for inverse problems: arXiv preprint arXiv:1805.11718.

Hale, D., 2009, Image-guided blended neighbor interpolation of scattered data: SEG Technical Program Expanded Abstracts 2009, Society of Exploration Geophysicists, 1127–1131.

Halliday, D., and A. Curtis, 2009, Generalized optical theorem for surface waves and layered media: Physical Review E, **79**, 056603.

Halliday, D., A. Curtis, J. Robertsson, and D.-J. van Manen, 2007, Interferometric surface-wave isolation and removal: Geophysics, **72**, A69–A73.

Halliday, D., A. Curtis, P. Vermeer, C. Strobbia, A. Glushchenko, D.-J. van Manen, and J. Robertsson, 2010, Interferometric ground-roll removal: Attenuation of scattered surface waves in single-sensor data: Geophysics, **75**, SA15–SA25.

Hansen, P. C., 1992, Analysis of Discrete Ill-Posed Problems by Means of the L-Curve: SIAM Review, **34**, 561–580.

———, 1998, Rank-deficient and discrete ill-posed problems : numerical aspects of linear inversion: Society for Industrial and Applied Mathematics.

Harper, F. M., and J. A. Konstan, 2016, The movielens datasets: History and context: Acm transactions on interactive intelligent systems (tiis), **5**, 19.

Hawkins, R., and M. Sambridge, 2015, Geophysical imaging using trans-dimensional trees: Geophysical Journal International, **203**, 972–1000.

Herceg, M., I. Artemieva, and H. Thybo, 2015, Sensitivity analysis of crustal correction for calculation of lithospheric mantle density from gravity data: Geophysical Journal International, **204**, 687–696.

Hicks, G. J., and R. G. Pratt, 2001, Reflection waveform inversion using local descent methods: Estimating attenuation and velocity over a gas-sand deposit – Waveform Inversion over a Gas Sand: Geophysics, **66**, 598–612.

Hinton, G. E., and R. R. Salakhutdinov, 2006, Reducing the dimensionality of data with neural networks: science, **313**, 504–507.

Hochreiter, S., and J. Schmidhuber, 1997, Long short-term memory: Neural computation, **9**, 1735–1780.

Huang, L., X. Dong, and T. E. Clee, 2017, A scalable deep learning platform for identifying geologic features from seismic attributes: The Leading Edge, **36**, 249–256.

Hunter, J., S. Pullan, R. Burns, R. Gagne, and R. Good, 1984, Shallow seismic reflection mapping of the overburden-bedrock interface with the engineering seismograph –some simple techniques: Geophysics, **49**, 1381–1385.

Ishii, M., and J. Tromp, 1999, Normal-mode and free-air gravity constraints on lateral variations in velocity and density of earth's mantle: Science, **285**, 1231–1236.

———, 2001, Even-degree lateral variations in the earth's mantle constrained by free oscillations and the free-air gravity anomaly: Geophysical Journal International, **145**, 77–96.

Jeong, W., H. Lee, and D. Min, 2012, Full waveform inversion strategy for density in the frequency domain: Geophysical Journal International, **188**, 1221–1242.

Kato, M., and H. Kawakatsu, 2001, Seismological in situ estimation of density jumps across the transition zone discontinuities beneath japan: Geophysical research letters, **28**, 2541–2544.

Käufl, P., A. Valentine, R. de Wit, and J. Trampert, 2015, Robust and fast probabilistic source parameter estimation from near-field displacement waveforms using pattern recognition: Bulletin of the Seismological Society of America, **105**, 2299–2312.

Käufl, P., A. P. Valentine, R. W. de Wit, and J. Trampert, 2016, Solving probabilistic inverse problems rapidly with prior samples: Geophysical Journal International, **205**, 1710–1728.

Käufl, P., A. P. Valentine, T. B. O'Toole, and J. Trampert, 2014, A framework for fast probabilistic centroid-moment-tensor determination – inversion of regional static displacement measurements: Geophysical Journal International, **196**, 1676–1693.

King, S., and A. Curtis, 2011, Velocity analysis using both reflections and refractions in seismic interferometry: Geophysics, **76**, SA83–SA96.

King, S., A. Curtis, and T. Poole, 2011, Interferometric velocity analysis using physical and nonphysical energy: Geophysics, **76**, SA35–SA49.

Kingma, D. P., and J. Ba, 2014, Adam: A method for stochastic optimization: arXiv preprint arXiv:1412.6980.

Kingma, D. P., and M. Welling, 2013, Auto-encoding variational bayes: arXiv preprint arXiv:1312.6114.

Kong, Q., D. T. Trugman, Z. E. Ross, M. J. Bianco, B. J. Meade, and P. Gerstoft, 2018, Machine learning in seismology: Turning data into insights: Seismological Research Letters, **90**, 3–14.

Laloy, E., R. Hérault, J. Lee, D. Jacques, and N. Linde, 2017, Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network: Advances in water resources, **110**, 387–405.

Langston, C. A., 2007a, Spatial gradient analysis for linear seismic arrays: Bulletin of the Seismological Society of America, **97**, 265–280.

———, 2007b, Wave gradiometry in the time domain: Bulletin of the Seismological Society of America, **97**, 926–933.

———, 2007c, Wave gradiometry in two dimensions: Bulletin of the Seismological Society of America, **97**, 401–416.

Langston, C. A., and M. M. Ayele, 2016, Vertical seismic wave gradiometry: Application at the san andreas fault observatory at depth: Geophysics, **81**, D233–D243.

Lapuschkin, S., A. Binder, G. Montavon, K.-R. Muller, and W. Samek, 2016, Analyzing classifiers: Fisher vectors and deep neural networks: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2912–2920.

Larochelle, H., and I. Murray, 2011, The neural autoregressive distribution estimator: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 29–37.

LeCun, Y., Y. Bengio, et al., 1995, Convolutional networks for images, speech, and time series: The handbook of brain theory and neural networks, **3361**, 1995.

LeCun, Y., L. Bottou, Y. Bengio, P. Haffner, et al., 1998, Gradient-based learning applied to document recognition: Proceedings of the IEEE, **86**, 2278–2324.

Leung, K., E. Schmerling, and M. Pavone, 2016, Distributional prediction of human driving behaviours using mixture density networks: Technical report, Technical report, Stanford University.

Li, H., W. Su, C.-Y. Wang, Z. Huang, and Z. Lv, 2010, Ambient noise love wave tomography in the eastern margin of the tibetan plateau: Tectonophysics, **491**, 194–204.

Li, Y., 2005, A study on applicability of density inversion in defining reservoirs: SEG Technical Program Expanded Abstracts 2005, Society of Exploration Geophysicists, 1646–1649.

Li, Y., Q. Guo, Z. Li, and T. Alkhalifah, 2019, Elastic reflection waveform inversion with variable density: Geophysics, **84**, R553–R567.

Liang, C., and C. A. Langston, 2009, Wave gradiometry for USArray: Rayleigh waves: Journal of Geophysical Research: Solid Earth, **114**.

Lin, F.-C., M. P. Moschetti, and M. H. Ritzwoller, 2008, Surface wave tomography of the western United States from ambient seismic noise: Rayleigh and Love wave phase velocity maps: Geophysical Journal International, **173**, 281–298.

Lin, F.-C., and M. H. Ritzwoller, 2011, Helmholtz surface wave tomography for isotropic and azimuthally anisotropic structure: Geophysical Journal International, **186**, 1104–1120.

Lin, F.-C., M. H. Ritzwoller, and R. Snieder, 2009, Eikonal tomography: surface wave tomography by phase front tracking across a regional broad-band seismic array: Geophysical Journal International, **177**, 1091–1110.

Liu, Y., and W. E. Holt, 2015, Wave gradiometry and its link with helmholtz equation solutions applied to USArray in the eastern US: Journal of Geophysical Research: Solid Earth, **120**, 5717–5746.

Liu, Y., M. Landrø, B. Arntsen, and J. van der Neut, 2017, A new approach to separate seismic time-lapse time shifts in the reservoir and overburden: Geophysics, **82**, 1–86.

Liu, Y., J. van der Neut, and B. Arntsen, 2016, Combination of surface and borehole seismic data for robust target-oriented imaging: Geophysical Journal International, **206**, 758–775.

Lobkis, O. I., and R. L. Weaver, 2001, On the emergence of the green's function in the correlations of a diffuse field: The Journal of the Acoustical Society of America, **110**, 3011–3017.

Lomas, A., and A. Curtis, 2018, 3D Marchenko Redatuming using 2D and 3D Seismic Data: Presented at the 80th EAGE Conference and Exhibition, Extended Abstracts, EAGE.

Lumley, D., 2010, 4D seismic monitoring of CO2 sequestration: The Leading Edge, **29**, 150–155.

Mairal, J., F. Bach, J. Ponce, et al., 2014, Sparse modeling for image and vision processing: Foundations and Trends® in Computer Graphics and Vision, **8**, 85–283.

Maiti, S., R. Krishna Tiwari, and H.-J. Kümpel, 2007, Neural network modelling and classification of lithofacies using well log data: a case study from KTB borehole site: Geophysical Journal International, **169**, 733–746.

Makansi, O., E. Ilg, O. Cicek, and T. Brox, 2019, Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7144–7153.

Martin, G. S., K. J. Marfurt, and S. Larsen, 2002, Marmousi-2: An updated model for the investigation of AVO in structurally complex areas, *in* SEG Technical Program Expanded Abstracts 2002: Society of Exploration Geophysicists, 1979–1982.

McCormack, M. D., D. E. Zaucha, and D. W. Dushek, 1993, First-break refraction event picking and seismic data trace editing using neural networks: Geophysics, **58**, 67–78.

McMechan, G. A., 1983, Seismic tomography in boreholes: Geophysical Journal International, **74**, 601–612.

Meier, U., A. Curtis, and J. Trampert, 2007a, Fully nonlinear inversion of fundamental mode surface waves for a global crustal model: Geophysical Research Letters, **34**.

———, 2007b, Global crustal thickness from neural network inversion of surface wave data: Geophysical Journal International, **169**, 706–722.

Meier, U., J. Trampert, and A. Curtis, 2009, Global variations of temperature and water content in the mantle transition zone from higher mode surface waves: Earth and Planetary Science Letters, **282**, 91–101.

Meles, G., K. Wapenaar, and A. Curtis, 2016, Reconstructing the primary reflections in seismic data by Marchenko redatuming and convolutional interferometry: Geophysics, **81**, Q15–Q26.

Meles, G. A., K. Löer, M. Ravasi, A. Curtis, and C. A. da Costa Filho, 2015, Internal multiple prediction and removal using Marchenko autofocusing and seismic interferometry: Geophysics, **80**, A7–A11.

Mildner, C., M. Dukalski, P. Elison, K. De Vos, F. Broggini, and J. Robertsson, 2019, True amplitude-versus-offset green's function retrieval using augmented Marchenko focusing: Presented at the 81st EAGE Conference and Exhibition 2019.

Moldoveanu, M., P. Caprioli, B. Kjellesvig, M. Ishak, M. Beecher, L. Mulidheva, and S. Pai, 2017, Marine seismic acquisition with autonomous marine vehicles towing 3D sensor arrays: The Leading Edge, **36**, 558–565.

Montagner, J.-P., and N. Jobert, 1988, Vectorial tomography–II. Application to the Indian Ocean: Geophysical Journal International, **94**, 309–344.

Montavon, G., S. Bach, A. Binder, W. Samek, and K.-R. Müller, 2017, Explaining nonlinear classification decisions with deep Taylor decomposition: Pattern Recognition, **65**, 211–222.

Montelli, R., G. Nolet, G. Masters, F. Dahlen, and S.-H. Hung, 2004, Global P and PP traveltime tomography: rays versus waves: Geophysical Journal International, **158**, 637–654.

Mordret, A., M. Landès, N. Shapiro, S. Singh, and P. Roux, 2014, Ambient noise surface wave tomography to determine the shallow shear velocity structure at Valhall: depth inversion with a neighbourhood algorithm: Geophysical Journal International, **198**, 1514–1525.

Mordret, A., N. M. Shapiro, S. Singh, P. Roux, J.-P. Montagner, and O. I. Barkved, 2013a, Azimuthal anisotropy at Valhall: The helmholtz equation approach: Geophysical Research Letters, **40**, 2636–2641.

Mordret, A., N. M. Shapiro, S. S. Singh, P. Roux, and O. I. Barkved, 2013b, Helmholtz tomography of ambient noise surface wave data to estimate Scholte wave phase velocity at Valhall life of the field: Geophysics, **78**, WA99–WA109.

Mosegaard, K., and A. Tarantola, 1995, Monte Carlo sampling of solutions to inverse problems: Journal of Geophysical Research: Solid Earth, **100**, 12431–12447.

Mosser, L., O. Dubrule, and M. J. Blunt, 2017, Reconstruction of three-dimensional porous media using generative adversarial neural networks: Physical Review E, **96**, 043309.

———, 2018, Stochastic seismic waveform inversion using generative adversarial networks as a geological prior: arXiv, arXiv:1806.03720.

Moya, A., and K. Irikura, 2010, Inversion of a velocity model using artificial neural networks: Computers and Ge, **36**, 1474–1483.

Muijs, R., O. J. Robertsson, A. Curtis, and K. Holliger, 2003, Near-surface seismic properties for elastic wavefield decomposition: Estimates based on multicomponent land and seabed recordings: Geophysics, **68**, 2073–2081.

Murat, M. E., and A. J. Rudman, 1992, Automated first arrival picking: A neural network approach: Geophysical Prospecting, **40**, 587–604.

Nawaz, M. A., and A. Curtis, 2017, Bayesian inversion of seismic attributes for geological facies using a hidden Markov model: Geophysical Journal International, **208**, 1184–1200.

———, 2018, Variational Bayesian inversion (VBI) of quasi-localized seismic attributes for the spatial distribution of geological facies: Geophysical Journal International, **214**, 845–875.

———, 2019, Rapid discriminative variational Bayesian inversion of geophysical data for the spatial distribution of geological properties: Journal of Geophysical Research: Solid Earth, 845–875.

Nicolson, H., A. Curtis, and B. Baptie, 2014, Rayleigh wave tomography of the British Isles from ambient seismic noise: Geophysical Journal International, **198**, 637–655.

Nicolson, H., A. Curtis, B. Baptie, and E. Galetti, 2012, Seismic interferometry and ambient noise tomography in the British Isles: Proceedings of the Geologists' Association, **123**, 74–86.

Niu, L., J. Ma, J. Geng, D. Zhou, and X. Yin, 2015, Mixture density network applied to reservoir parameter inversion in bohai oil field, *in* SEG Technical Program Expanded Abstracts 2015: Society of Exploration Geophysicists, 2806–2810.

Operto, S., Y. Gholami, V. Prieux, A. Ribodetti, R. Brossier, L. Metivier, and J. Virieux, 2013, A guided tour of multiparameter full-waveform inversion with multicomponent data: From theory to practice: The Leading Edge, **32**, 1040–1054.

Ormoneit, D., and R. Neuneier, 1996, Experiments in predicting the german stock index dax with density estimating neural networks: IEEE/IAFE 1996 Conference on Computational Intelligence for Financial Engineering (CIFEr), IEEE, 66–71.

Phillips, P. J., H. Wechsler, J. Huang, and P. J. Rauss, 1998, The FERet database and evaluation procedure for face-recognition algorithms: Image and vision computing, **16**, 295–306.

Piana Agostinetti, N., G. Giacomuzzi, and A. Malinverno, 2015, Local three-dimensional earthquake tomography by trans-dimensional Monte Carlo sampling: Geophysical Journal International, **201**, 1598–1617.

Poppeliers, C., P. Punoševac, and T. Bell, 2013, Three-dimensional seismic-wave gradiometry for scalar waves: Bulletin of the Seismological Society of America, **103**, 2151–2160.

Poulton, M. M., 2002, Neural networks as an intelligence amplification tool: A review of applications: Geophysics, **67**, 979–993.

Pratt, R., Z.-M. Song, P. Williamson, and M. Warner, 1996, Two-dimensional velocity models from wide-angle seismic data by wavefield inversion: Geophysical Journal International, **124**, 323–340.

Pratt, R. G., 1999, Seismic waveform inversion in the frequency domain, part 1: Theory and verification in a physical scale model: Geophysics, **64**, 888–901.

Prieux, V., R. Brossier, and S. Operto, 2013, Multiparameter full waveform inversion of multicomponent ocean-bottom-cable data from the Valhall field. Part 1: Imaging compressional wave speed, density and attenuation: Geophysical Journal International, **194**, 1640–1664.

Ravasi, M., 2017, Rayleigh-Marchenko redatuming for target-oriented, true-amplitude imaging: Geophysics, **82**, S439–S452.

Ravasi, M., I. Vasconcelos, A. Kritski, A. Curtis, and C. A. da Costa Filho, 2016, Target-oriented Marchenko imaging of a North Sea field: Geophysical Journal International, **205**, 99–104.

Rawlinson, N., A. Fichtner, M. Sambridge, and M. K. Young, 2014, Seismic tomography and the assessment of uncertainty, *in* Advances in Geophysics: Elsevier, **55**, 1–76.

Rawlinson, N., S. Pozgay, and S. Fishwick, 2010, Seismic tomography: a window into deep earth: Physics of the Earth and Planetary Interiors, **178**, 101–135.

Rawlinson, N., and M. Sambridge, 2004, Wave front evolution in strongly heterogeneous layered media using the Fast Marching Method: Geophysical Journal International, **156**, 631–647.

———, 2005, The Fast Marching Method: An effective tool for tomographic imaging and tracking multiple phases in complex layered media: Exploration Geophysics, **36**, 341–350.

Ritzwoller, M. H., and A. L. Levshin, 1998, Eurasian surface wave tomography: Group velocities: Journal of Geophysical Research: Solid Earth, **103**, 4839–4878.

Ritzwoller, M. H., N. M. Shapiro, M. P. Barmin, and A. L. Levshin, 2002, Global surface wave diffraction tomography: Journal of Geophysical Research: Solid Earth, **107**, ESE–4.

Romanowicz, B., 1995, A global tomographic model of shear attenuation in the upper mantle: Journal of Geophysical Research: Solid Earth, **100**, 12375–12394.

Root, B., J. Ebbing, W. van der Wal, R. England, and L. Vermeersen, 2016, Comparing gravity-based to seismic-derived lithosphere densities: a case study of the british isles and surrounding areas: Geophysical Journal International, **208**, 1796–1810.

Rose, J., 2002, Single-sided autofocusing of sound in layered materials: Inverse Problems, **18**, 1923.

Roth, G., and A. Tarantola, 1994, Neural networks and inversion of seismic data: Journal of Geophysical Research, **99**, 6753–6768.

Roux, P., 2009, Passive seismic imaging with directive ambient noise: application to surface waves and the san andreas fault in parkfield, ca: Geophysical Journal International, **179**, 367–373.

Ruigrok, E., D. Draganov, and K. Wapenaar, 2008, Global-scale seismic interferometry: theory and numerical examples: Geophysical Prospecting, **56**, 395–417.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams, 1985, Learning internal representations by error propagation: Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Sabra, K. G., P. Gerstoft, P. Roux, W. Kuperman, and M. C. Fehler, 2005, Surface wave tomography from microseisms in southern california: Geophysical Research Letters, **32**.

Saito, M., 1988, DISPER80 : A subroutine package for the calculation of seismic normal mode solutions: Academic Press.

Sambridge, M., 1999, Geophysical inversion with a neighbourhood algorithm - II. Appraising the ensemble: Geophysical Journal International, **138**, 727–746.

Sambridge, M., and K. Mosegaard, 2002, Monte Carlo methods in geophysical inverse problems: Reviews of Geophysics, **40**, 3–1.

Schmelzbach, C., S. Donner, H. Igel, D. Sollberger, T. Taufiqurrahman, F. Bernauer, M. Häusler, C. Van Renterghem, J. Wassermann, and J. Robertsson, 2018, Advances in 6-C seismology: applications of combined translational and rotational motion measurements in global and exploration seismology: Geophysics, **83**, 1–58.

Schütt, K. T., F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, 2017, Quantum-chemical insights from deep tensor neural networks: Nature communications, **8**, 13890.

Shahraeeni, M. S., and A. Curtis, 2011, Fast probabilistic nonlinear petrophysical inversion: Geophysics, **76**, E45–E58.

Shahraeeni, M. S., A. Curtis, and G. Chao, 2012, Fast probabilistic petrophysical mapping of reservoirs from 3D seismic data: Geophysics, **77**, O1–O19.

Shapiro, N., and M. Ritzwoller, 2002, Monte-Carlo inversion for a global shear-velocity model of the crust and upper mantle: Geophysical Journal International, **151**, 88–105.

Shapiro, N. M., and M. Campillo, 2004, Emergence of broadband Rayleigh waves from correlations of the ambient seismic noise: Geophysical Research Letters, **31**.

Shapiro, N. M., M. Campillo, L. Stehly, and M. H. Ritzwoller, 2005, High-resolution surface-wave tomography from ambient seismic noise: Science, **307**, 1615–1618.

Simons, F. J., R. D. Van Der Hilst, J.-P. Montagner, and A. Zielhuis, 2002, Multimode Rayleigh wave inversion for heterogeneity and azimuthal anisotropy of the Australian upper mantle: Geophysical Journal International, **151**, 738–754.

Simons, F. J., A. Zielhuis, and R. D. Van Der Hilst, 1999, The deep structure of the Australian continent from surface wave tomography: Lithos, **48**, 17–43.

Singer, J., A. Obermann, E. Kissling, H. Fang, G. Hetényi, and D. Grujic, 2017, Along-strike variations in the h imalayan orogenic wedge structure in b hutan from ambient seismic noise tomography: Geochemistry, Geophysics, Geosystems, **18**, 1483–1498.

Singh, S., and R. Snieder, 2017, Strategies for imaging with Marchenko-retreived Green's functions: Geophysics, **82**, Q23–Q37.

Singh, S., R. Snieder, J. Behura, and J. van der Neut, 2015, Marchenko imaging: Imaging with primaries, internal multiples, and free-surface multiples: Geophysics, **80**, S164–S174.

Slob, E., and K. Wapenaar, 2017, Theory for Marchenko imaging of marine seismic data with free surface multiple elimination: Presented at the 79th EAGE Conference and Exhibition 2017.

Slob, E., K. Wapenaar, F. Broggini, and R. Snieder, 2014, Seismic reflector imaging using internal multiples with Marchenko-type equations: Geophysics, **79**, S63–S76.

Snieder, R., 2004, Extracting the green's function from the correlation of coda waves: A derivation based on stationary phase: Physical Review E, **69**, 046610.

Socco, L. V., S. Foti, and D. Boiero, 2010, Surface-wave analysis for building near-surface velocity models—established approaches and new perspectives: Geophysics, **75**, 75A83–75A102.

Sollberger, D., C. Schmelzbach, J. O. Robertsson, S. A. Greenhalgh, Y. Nakamura, and A. Khan, 2016, The shallow elastic structure of the lunar crust: New insights from seismic wavefield gradient analysis: Geophysical Research Letters, **43**, 10–078.

Staring, M., R. Pereira, H. Douma, J. van der Neut, and C. Wapenaar, 2017, Adaptive double-focusing method for source-receiver Marchenko redatuming on field data: Presented at the SEG Technical Program Expanded Abstracts 2017.

Stewart, P., 2006, Interferometric imaging of ocean bottom noise, *in* SEG Technical Program Expanded Abstracts 2006: Society of Exploration Geophysicists, 1555–1559.

Su, W.-j., R. L. Woodward, and A. M. Dziewonski, 1994, Degree 12 model of shear velocity hetereogeneity in the mantle: Journal of Geophysical Research: Solid Earth, **99**, 2156–6980.

Tanimoto, T., 1991, Waveform inversion for three-dimensional density and s wave structure: Journal of Geophysical Research: Solid Earth, **96**, 8167–8189.

Tarantola, A., 2005, Inverse problem theory: Siam.

Thompson, M., M. Andersen, R. Elde, S. Roy, and S. Skogland, 2015, The startup of permanent reservoir monitoring for Snorre and Grane: Presented at the 77th EAGE Conference and Exhibition 2015.

Trampert, J., and J. H. Woodhouse, 1995, Global phase velocity maps of Love and Rayleigh waves between 40 and 150 seconds: Geophysical Journal International, **122**, 675–690.

Uria, B., I. Murray, and H. Larochelle, 2013, RNADE: The real-valued neural autoregressive density-estimator: Advances in Neural Information Processing Systems, 2175–2183.

Valentine, A., and L. M. Kalnins, 2016, An introduction to learning algorithms and potential applications in geomorphometry and earth surface dynamics.: Earth surface dynamics., **4**, 445–460.

Valentine, A. P., and J. Trampert, 2012, Data space reduction, quality assessment and searching of seismograms: autoencoder networks for waveform data: Geophysical Journal International, **189**, 1183–1202.

van den Oord, A., S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, 2016a, Wavenet: A generative model for raw audio: arXiv preprint arXiv:1609.03499.

van den Oord, A., N. Kalchbrenner, and K. Kavukcuoglu, 2016b, Pixel recurrent neural networks: arXiv preprint arXiv:1601.06759.

Van der Baan, M., and C. Jutten, 2000, Neural networks in geophysical applications: Geophysics, **65**, 1032–1047.

van der Hilst, R. D., M. V. de Hoop, P. Wang, P. Shim, S.-H. amd Ma, and L. Tenorio, 2007, Seismostratigraphy and Thermal Structure of Earth's Core-Mantle Boundary Region: Science, **315**, 1813–1817.

Van der Lee, S., and G. Nolet, 1997, Upper mantle S velocity structure of North America: Journal of Geophysical Research: Solid Earth, **102**, 22815–22838.

van der Neut, J., J. Brackenhoff, M. Staring, L. Zhang, S. de Ridder, E. Slob, and K. Wapenaar, 2018, Single-and double-sided Marchenko imaging conditions in acoustic media: IEEE Transactions on Computational Imaging, **4**, 160–171.

van der Neut, J., J. Thorbecke, K. Mehta, E. Slob, and K. Wapenaar, 2011, Controlled-source interferometric redatuming by crosscorrelation and multidimensional deconvolution in elastic media: Geophysics, **76**, SA63–SA76.

van der Neut, J., K. Wapenaar, J. Thorbecke, and E. Slob, 2015, An illustration of adaptive Marchenko imaging: Leading Edge, **34**, 818–822.

van Manen, D.-J., A. Curtis, and J. Robertsson, 2006, Interferometric modeling of wave propagation in inhomogeneous elastic media using time reversal and reciprocity: Geophysics, **71**, SI47–SI60.

van Manen, D.-J., J. Robertsson, and A. Curtis, 2005, Modeling of Wave Propagation in Inhomogeneous Media: Physical Review Letters, **94**, 164301.

Vasconcelos, I., K. Wapenaar, J. van der Neut, C. Thomson, and M. Ravasi, 2015, Using Inverse Transmission Matrices for Marchenko redatuming in highly complex media: Presented at the 2015 SEG Annual Meeting, Society of Exploration Geophysicists.

Versteegh, M., R. Thiolliere, T. Schatz, X. N. Cao, X. Anguera, A. Jansen, and E. Dupoux, 2015, The zero resource speech challenge 2015: Presented at the Sixteenth Annual Conference of the International Speech Communication Association.

Villasenor, A., M. Ritzwoller, A. Levshin, M. Barmin, E. Engdahl, W. Spakman, and J. Trampert, 2001, Shear velocity structure of central Eurasia from inversion of surface wave velocities: Physics of the Earth and Planetary Interiors, **123**, 169–184.

Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74**, WCC1–WCC26.

Walker, M., and A. Curtis, 2014a, Expert elicitation of geological spatial statistics using genetic algorithms: Geophys. J. Int, **198**, 342–356.

———, 2014b, Varying prior information in Bayesian inversion: Inverse Problems, **30**, 065002.

Wang, W., S. Xu, and B. Xu, 2016, Gating recurrent mixture density networks for acoustic modeling in statistical parametric speech synthesis: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 5520–5524.

Wang, Y., and R. G. Pratt, 1997, Sensitivities of seismic traveltimes and amplitudes in reflection tomography: Geophysical journal international, **131**, 618–642.

Wang, Y., R. E. White, and R. G. Pratt, 2000, Seismic amplitude inversion for interface geometry: practical approach for application: Geophysical Journal International, **142**, 162–172.

Wang, Z., A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al., 2004, Image quality assessment: from error visibility to structural similarity: IEEE transactions on image processing, **13**, 600–612.

Wapenaar, K., 2004, Retrieving the Elastodynamic Green's Function of an Arbitrary Inhomogeneous Medium by Cross Correlation: Physcial Review Letters, **93**, 254301.

———, 2014, Single-sided Marchenko focusing of compressional and shear waves: Physical Review E, **90**, 063202.

Wapenaar, K., and J. Fokkema, 2006, Green's function representations for seismic interferometry: Geophysics, **71**, SI33–SI46.

Wapenaar, K., J. Thorbecke, and J. van der Neut, 2014, Marchenko imaging: Geophysics, **79**, WA39–WA57.

Wapenaar, K., J. van der Neut, and E. Ruigrok, 2011, Seismic interferometry by crosscorrelation and by multidimensional deconvolution: a systematic comparison: Geophysical Journal International, **185**, 1335–1364.

Williams, P. M., 1996, Using neural networks to model conditional multivariate densities: Neural Computation, **8**, 843–854.

Woodhouse, J. H., and A. M. Dziewonski, 1984, Mapping the upper mantle: Three-dimensional modeling of Earth structure by inversion of seismic waveforms: Journal of Geophysical Reseach: Solid Earth, **89**, 5953–5986.

Xia, J., R. D. Miller, and C. B. Park, 1999, Estimation of near-surface shear-wave velocity by inversion of rayleigh waves: Geophysics, **64**, 691–700.

Xiao, X., M. Zhou, and G. T. Schuster, 2006, Salt-flank delineation by interferometric imaging of transmitted P- to S-waves: Geophysics, **71**, SI197–SI207.

Yao, H., R. D. van Der Hilst, and M. V. De Hoop, 2006, Surface-wave array tomography in SE Tibet from ambient seismic noise and two-station analysis—I. Phase velocity maps: Geophysical Journal International, **166**, 732–744.

Zen, H., and A. Senior, 2014, Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 3844–3848.

Zhan, Z., Q. Li, and J. Huang, 2018, Application of wavefield compressive sensing in surface wave tomography: Geophysical Journal International, **213**, 1731–1743.

Zhang, X., A. Curtis, E. Galetti, and S. de Ridder, 2018, 3-D Monte Carlo surface wave tomography: Geophysical Journal International, **215**, 1644–1658.

Zhang, X., F. Hansteen, and A. Curtis, 2019, Fully 3D Monte Carlo ambient noise tomography over Grane field: Presented at the 81st EAGE Conference and Exhibition 2019.

Zheng, D., E. Saygin, P. Cummins, Z. Ge, Z. Min, A. Cipta, and R. Yang, 2017, Transdimensional Bayesian seismic ambient noise tomography across SE Tibet: Journal of Asian Earth Sciences, **134**, 86–93.

Zheng, S., X. Sun, X. Song, Y. Yang, and M. H. Ritzwoller, 2008, Surface wave tomography of china from ambient seismic noise correlation: Geochemistry, Geophysics, Geosystems, **9**.

Zhou, Y., G. Nolet, F. Dahlen, and G. Laske, 2006, Global upper-mantle structure from finite-frequency surface-wave tomography: Journal of Geophysical Research: Solid Earth, **111**.